IEEE
WIRELESS COMMUNICATIONS

December 2002 Vol. 9 No. 6

- Co-Channel Interference Cancellation
- Multi-Antenna Wireless Systems
- Vertical Optimization of Data Transmission
- Stanford Interactive Workspaces
- Facilitating the Programming of the Smart Home
- Prediction Algorithms in the MavHome Architecture

Wireless Local Area Networks

Designing and Implementing Smart Homes

A Publication of the IEEE Communications Society
In cooperation with IEEE Computer and Vehicular Technology Societies

IEEE

We have low power analog ICs for your multi-channel fiber optic transceivers.

Introducing RF Micro Devices' standard and custom TIAs, limiting amplifiers and VCSEL/Edge Emitting laser drivers with unique features in four-channel configurations. RFMD® engineers utilize Optimum Technology Matching® to design premium analog ASICs through 12 Gbps for your multi-channel fiber optic modules.

If your goal is to design fiber optic transceivers optimized for low power and high data throughput, then you need RFMD's low-cost analog circuits.

RX

RF3734
Quad Limiting Amplifier Serial at 3.125 Gbps

- 430mW power dissipation at +3.3V supply
- Minimum output signal 400mV differential (200mV SE)
- 4GHz bandwidth
- Wide limiting range (10/500mV)
- 15 ps p-p maximum jitter generation at 25mV input
- 50 ps typical rise/fall time
- Input/output return loss <10 dB
- Programmable level for loss of signal alarm function

RF3744
Quad Channel Receiver (TIA+LA+LOS) at 3.125 Gbps

- 860mW power dissipation at +3.3V supply
- 2.5GHz bandwidth
- Optical sensitivity –20 dBm
- Optical overload 0 dBm
- -40 ps p-p typical jitter generation at 25mV input
- 2000 V/A trans-impedance
- 50 ohms output impedance

TX

RF3754 and RF3764 Laser Drivers

- +3.3V supply voltage, 36mA quiescent current
- Common cathode driver
- Automatic laser power control/constant output power mode
- Integrated safety circuits
- Power-on reset signal

RF3754
Quad Channel VCSEL Laser Driver at 3.125 Gbps

- 1-18mA laser modulation current
- 2-25mA laser bias current control range

RF3764
Quad Channel Edge Emitting Laser Driver at 3.125 Gbps

- 3-50mA laser modulation current
- 2-50mA laser bias current control range

Excellent service and support is our top priority.

For sales or technical support, contact **336.678.5570** or **callcenter@rfmd.com.**

RF MICRO·DEVICES
ProvidingCommunicationSolutions™

Handsets WLAN *Bluetooth*™ GPS **Optical Communications** Infrastructure Broadband

www.rfmd.com

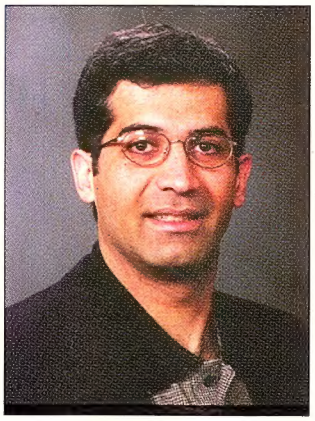
ISO 9001: 2000 Certified

DECEMBER 2002/VOL. 9 NO. 6

IEEE WIRELESS COMMUNICATIONS

Cover illustration: Corbis

WIRELESS LANS AND SMART HOMES



MAHMOUD NAGHSHINEH

Welcome to our December issue of *IEEE Wireless Communications*. This issue covers wireless LANs and smart homes, two topics in our field receiving a great amount of attention. Wireless LANs have emerged as a mainstream networking technology in enterprises, homes, and recently in public areas. In late '80s and early '90s, wireless LANs were considered mostly experimental technologies looking for markets and application areas. Today, in contrast to a decade ago, they are viewed as an alternative to 3G data networking and drive considerable business opportunities in a number of markets. Several events and emerging technologies have contributed to this. First of all, standards have matured, and there has been increased focus on interoperability. Here IEEE 802.11 and lately 802.15 working groups have especially contributed to progress. Moreover, many networking products and services have paid a lot of attention to internetworking of wireless LANs and wired (mostly Ethernet) LANs, as well as solving the immediate security and network management problems. More important, the cost of technology has decreased about an order of magnitude in a very similar fashion to the affordability and cost of technology reduction that helped Ethernet adoption in the past. As a result, enterprise chief information officers find wireless LANs offering a strong value proposition as far as return on investment in usability, reduction of wiring costs, and operations management. This has given rise to accelerated adoption of wireless LANs in enterprises and office environments. The other very important factor in adoption of wireless technologies has been the proliferation of Internet access. This has been a key driver

IEEE Wireless Communications (ISSN 1536-1284) is published bimonthly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters address: IEEE, 3 Park Avenue, 17th Floor, New York, NY 10016-5997; tel: 212-705-8900; fax: 212-705-8999; e-mail: c.kemelmacher@comsoc.org. Responsibility for the contents rests upon authors of signed articles and not the IEEE or its members. Unless otherwise specified, the IEEE neither endorses nor sanctions any positions or actions espoused in *IEEE Wireless Communications*.

Annual subscription: Member subscription: $25 per year; Non-member subscription: $200 per year. Single copy: $10 for members and $20 for nonmembers.

Editorial correspondence: Manuscripts for consideration may be submitted to the Editor-in-Chief: Mahmoud Nagshineh, IBM Watson Research Center, 30 Saw Mill River Road, Hawthorne, NY 10532. Electronic submissions may be sent in postscript to: mahmoud@watson.ibm.com.

Postmaster: Send address changes to *IEEE Wireless Communications*, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331; or email to address.change@ieee.org. Printed in USA. Periodicals postage paid at New York, NY and at additional mailing offices. Canadian GST #40012498.

Subscriptions: Send orders, address changes to: IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08855-1331; tel.: 908-981-0060.

Advertising: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, *IEEE Wireless Communications*, 305 East 47th Street, New York, NY 10017-2394.

of wireless technologies, mostly in consumer spaces. In addition, emerging VPN standards and secure tunnel protocols, and accelerated access availability and affordability from Internet service providers have simplified wireless access by enterprise workers, their customers, and consumers. We also need to keep in mind that there is has been a lot of support from the mainstream operating systems, and networking middleware in mobile and portable devices to support easy access to wireless networks. As a result, wireless access in a true sense of anywhere, anytime has become a reality and a mainstream paradigm for access in enterprises, public areas, and homes.

There is a lot of progress in this area. But this is not the end, but rather really the beginning of a mainstream technology. There are many efforts to offer higher speeds in higher bands, and many standards underway to pave the path for such higher speeds and new applications such as packetized voice and multimedia. Another very important standard in the wireless access field is its extension into personal area networking (PAN). Through Bluetooth Special Interest Group (SIG) and IEEE 802.15 we are witnessing widespread support for wireless PANs. This is mainly driven by smart phones and PDAs. While there is much debate in industry and academia about positioning of Bluetooth and wireless LANs, I think we need to look at wireless PANs and LANs as two sides of the same coin. Both technologies have strong value propositions and need to be focused on their very different application scenarios. The whole wireless space (PANs, LANs, and cellular or WANs) suffers from the fact that there are a number of diverse access technologies with overlapping application scenarios. Through development of multiband access technologies, reconfigurable radio systems, and other innovative methods, we must reach a point where the specificity of wireless access is hidden from users in a similar way to how modem access over telephone lines works today.

Given the above-mentioned progress in affordability, usability, and proliferation of the infrastructure, and supporting features in device hardware and software, we are witnessing an expansion of wireless technologies and applications in smart homes. This expansion is really happening on two fronts. First, the access technology is being expanded into a number of other technologies such as phone and power lines. Second, there are efforts to extend the usage model beyond "traditional" mobile computing devices and phones into consumer electronics and many other elements of the home environment. In the smart home field, there are numerous research projects, a few commercial offerings, and a number of industry standards in place to define a framework for carrier technologies, media access methods, interoperability, lightweight Internet protocols, and new distributed applications. This expansion is certainly far broader and more complicated than what happened in the wireless LAN space. On one hand, this is simpler since there is already an established framework based on wireless LANs and PANs to build on. On the other hand, many new standards need to be defined, technologies need to reach maturity and affordability, and operating systems, embedded software, and applications need to provide the necessary support and usability for smart homes to become mainstream. In my opinion, the driving applications and usability issues need to be focused on and staged in order to build momentum behind smart home technologies.

This issue of *Wireless Communications* is organized in two sections. The Wireless LANs section is organized by Benny Bing, Chris Heegard, and Bob Heile. The Smart Homes section is organized by Sajal Das and Diane Cook. You will find a number of papers that cover key technologies, protocols, applications, and infrastructure topics. They address many issues discussed above and offer different viewpoints on important topics and the future. I would like to thank our guest editors for their great contribution and effort in putting this issue together. Also many thanks to many authors who submitted their papers in response to our calls for papers.

This issue is also my last issue as Editor-in-Chief. Words fall short in describing my gratitude to many who served as guest editors, and technical editors, authors, and editorial staff and production at IEEE Communications Society. This has been one of the most gratifying team experiences of my research career. The last three years have been great for me to serve as Editor-in-Chief of this magazine. It has been a tremendous opportunity to serve our magazine readers and the entire community by fostering presentation and publication of the latest research ideas, many detailed surveys and overviews, as well as numerous papers on architectures, technologies, applications, and services in the wireless area. I am very humbled and honored by all the support and encouragement I have been given. I am especially grateful to Mischa Schwartz, Anthony Acampora, past Editors-in-Chief Hamid Ahmadi and Tom LaPorta, past and present editorial board directors, especially Mark Karol and Jim Kurose, IEEE Communications Society staff, especially Sue Lange, Cathy Kemelmacher, Joseph Milizzo, and Jack Howell, and finally, senior leadership and many colleagues at IBM Research.

Now it is time to hand off to Michele Zorzi, Editor-in-Chief starting in 2003. I have great confidence in Michele and I know he will be great in this position. He has been with our magazine as a technical editor since 1996, and has served as associate EIC for the past year. As we "wireless people" know, it had better be a soft handoff or you lose the connection! We have a very soft handoff prepared that will preserve the continuity of the magazine and its tradition, and grow it into new dimensions. Michele will take it from here.

Have a happy and healthy season!

EDITED BY YUGUANG FANG

▲Mobility Increases the Capacity of Ad Hoc Wireless Networks

M. Grossglauser and D.N.C. Tse, IEEE/ACM Transactions on Networking, 10(4), 477-486, August 2002.

A fundamental characteristic of mobile wireless networks is the time variation of the channel strength of the underlying communication links. The impact of such time variation on the design of wireless networks permeates throughout the layers, ranging from coding and power control at the physical layer to cellular handoff and coverage planning at the network layer. To cope with the time variation, an important means is the use of diversity in time, frequency, and space. Clever deployment of diversity in mobile ad hoc networks can enhance the transmission quality, hence the capacity. This paper focuses on the study of the capacity of ad hoc wireless networks, which is constrained by the mutual interference of concurrent transmissions between nodes. The authors study a model of an ad hoc network where n mobile nodes communicate in random source-dSestination pairs and examine the per-session throughput for applications with loose delay constraints, such that the topology changes over the timescale of packet delivery. Under this assumption, the per-user throughput is shown to increase dramatically when nodes are mobile rather than fixed. This improvement is achieved by exploiting a form of multiuser diversity via packet relaying. This paper complements well the paper by Gupta and Kumar ("The Capacity of Wireless Networks," *IEEE Transactions on Information Theory*, 46, 388-404, March 2000), where all nodes are assumed to be fixed.

▲ SPINS: Security Protocols for Sensor Networks

A. Perrig, R. Szewczyk, J.D. Tygar, V. Wen, and D.E. Culler, ACM Wireless Networks, 8(5), 521-534, September 2002.

Wireless sensor networks have received intensive attention in the last few years; it is envisioned that self-organizing sensors will be widely deployed in the near future. While much research has focused on making these networks feasible and useful, security has received little attention. In this paper, the authors present a suite of security protocols that are optimized for sensor networks: SPINS. SPINS has two secure building blocks: SNEP (Secure Network Encryption Protocol) and mTESLA (the "micro" version of TESLA). SNEP provides data confidentiality, two-party data authentication, and evidence of data freshness with low overhead, while mTESLA provides authenticated broadcast for severely resource-constrained environments. The above protocols have been implemented and shown to be practical even on minimal hardware: the performance of the protocol suite easily matches the data rate of the sensor networks, and the protocol suite can be used for building higher-level protocols.

▲ A Composable Framework for Secure Multi-Modal Access to Internet Services from Post-PC Devices

J. Ross, J. L. Hill, M. Y. Chen, A. D. Joseph, D. E. Culler, and E. A. Brewer, Mobile Networks and Applications, 7(5), 389-406, October 2002.

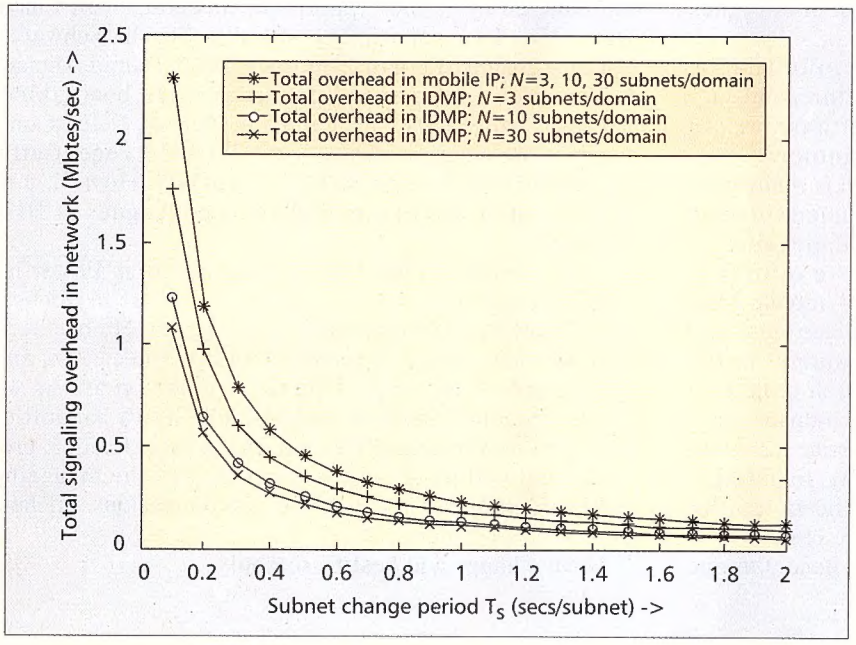
The post-PC revolution is bringing information access to a wide range of devices beyond the desktop, such as public kiosks, and mobile devices like cellular telephones, PDAs, and voice-based vehicle telematics. However, existing deployed Internet services are geared toward the secure rich interface of private desktop computers. In this paper, the authors propose the use of an infrastructure-based secure proxy architecture to bridge the gap between the capabilities of post-PC devices and the requirements of Internet services. By combining generic content and security transformation functions with service-specific rules, the architecture decouples device capabilities from service requirements, and simplifies the addition of new devices and services. Security and protocol specifics are abstracted into reusable components. Moreover, the architecture offers the ability to deal with untrusted public Internet access points by fine-grained control over the content and functionality exposed to the esnd device, as well as support for using trusted and untrusted devices in tandem.

▲Energy-Efficient Broadcast and Multicast Trees in Wireless Networks

J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, ACM Mobile Networks and Applications, 7(6), 481-492, December 2002.

The wireless networking environment presents formidable challenges to the study of broadcasting and multicasting problems. In this paper, the authors focus on the problem of multicast tree construction, and introduce and evaluate algorithms for tree construction in infras-

ERRATA



In the June issue of IEEE Wireless Communications, in the article entitled "IDMP: An Intradomain Mobility Management Protocol for Next Generation Wireless Networks," Figure 10 was published incorrectly. The correct figure is published here.

tructureless all-wireless applications. The performance metric used to evaluate broadcast and multicast trees is energy efficiency, and the Broadcast Incremental Power (BIP) algorithm is developed and adapted to multicast operation by introducing the Multicast Incremental Power (MIP) algorithm. These algorithms exploit the broadcast nature of wireless communication environments, and address the need for energy-efficient operation. It has been shown that these algorithms provide better performance than algorithms that have been developed for link-based wired environments.

▲Implementing Automatic Location Update for Follow-Me Database Using VoIP and Bluetooth Technologies

Y. B. Lin, H. Y. Cheng, Y. H. Cheng, and P. Agrawal, IEEE Transactions on Computers, Special Issue on Data Management Systems and Mobile Computing, 51(10), 1154-1168, October 2002.

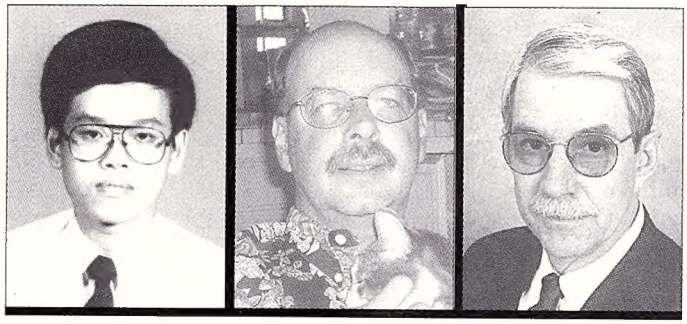
Personal Number (PN) service or Follow-Me service allows a user to access telecommunications services with any terminal in any location within the service area. To provide this feature, the PN user needs to manually register with a phone number every time he/she enters a new location. If a user forgets to register the new phone number, the incoming calls will be misrouted. To provide user-friendly PN service, the authors proposes an Automatic Follow-me Service (AFS) that automatically updates the PN records in the Follow-me database. The proposed scheme can easily be integrated with the existing Follow-me database to automate the PN services offered by different service providers. The authors show how the AFS is implemented by using voice over IP (VoIP) and Bluetooth technologies. Detailed performance evaluation is carried out through an analytical approach as well as simulation, and demonstrates how the design parameter polling frequency is chosen to optimize the AFS performance.

▲General Modeling and Performance Analysis for Location Management in Wireless Mobile Networks

Y. Fang, IEEE Transactions on Computers, Special Issue on Data Management Systems and Mobile Computing, 51(10), 1169-1181, October 2002.

Location management plays a significant role in the current and the future wireless mobile networks in effectively delivering services to the mobile users on the move. Many schemes have been proposed and investigated extensively in the past. However, most performance analyses were carried out either under simplistic assumptions on some time variables or via simulations. In this paper, the author presents a new analytical approach to investigate the trade-off (cost) analysis for location management schemes under fairly general assumption (all time variables are assumed to be generally distributed). The author focuses on two specific location management schemes, Pointer Forwarding Scheme (PFS) and Two-Location Algorithm (TLA), and presents analytical formulae for the total costs, which are easy to compute. Numerical results show that the traditional approximate exponential modeling for some time variables may lead to wrong decisions in the trade-off analysis while the analytical results presented in this paper can easily be applied to obtain the correct decision.

WIRELESS LANS



BENNY BING CHRIS HEEGARD BOB HEILE

Wireless local area networks emerged from near disaster in the mid-'90s to become one of today's driving wireless technologies, delivering high data rates on unlicensed spectrum for both the enterprise and the home.

The fact that U.S. megastores such as Starbucks and Sears are deploying wireless LANs shows how prevalent the technology has now become. However, there are several major wireless LAN standards (802.11a, 802.11b, and most recently 802.11g) operating on different frequency bands (2.4 and 5 GHz), and it is unclear which technology will eventually prevail, or whether combo solutions involving combined standards are required. With new wireless personal devices involving Bluetooth starting to gain traction commercially, the wireless world continues to create difficult options for both end users and service providers.

Wireless LANs provide high-speed, cable-free access primarily for indoor environments. A substantial portion of the cost of LAN deployment is in interconnecting end-user devices, which many networking experts acknowledge can sometimes exceed the cost of computer hardware and software. A wireless LAN removes the labor and material costs inherent in wiring. It also offers the flexibility to reconfigure or add more nodes to the network without much planning effort and the cost of recabling, thereby making future upgrades inexpensive and easy. The ability to add new mobile computing devices quickly is another main consideration for choosing a wireless LAN. Thus, the proliferation of cheaper, smaller, and more powerful portable notebook computers has fueled tremendous growth in the wireless LAN industry in recent years.

A wireless LAN need not transfer purely data traffic. It can also support packetized voice transmission. People today are spending huge amounts of money, even from office to office, calling on cellular phones. With a wireless LAN infrastructure, it costs them much less than it does using cellular phones or any other equipment. A more compelling use of wireless LANs is in overcoming the inherent limitations of wireless WANs. Current 3G wireless data rates go up to 2 Mb/s only with restricted mobility, whereas wireless LANs offer data rates of up to 54 Mb/s and operate on unlicensed frequency bands. This has led to some technologists predicting that eventually we are more likely to see dense urban broadband wireless LANs that are linked together into one network than widespread use of high-powered WAN handsets cramming many bits into expensive and narrow slices of radio spectrum.

The articles in this special issue focus on the state-of-the-art technologies in wireless LAN research. They cover a very broad scope, including all key layers of the network hierarchy. The first two articles deal with physical layer research. The article "Co-Channel Interference Cancellation Based on MIMO OFDM Systems" by Luca Giangaspero *et al.* demonstrates how multiple input multiple output (MIMO) systems can be exploited for broadband wireless indoor applications. Two MIMO OFDM-based systems, Wind-Flex and Ubiquitous Antenna, are considered. The authors aim to increase the system capacity of both systems using different approaches: increasing the single link data rate and increasing the number of users in the entire system. The next article, "Reduced Dimension Space-Time Processing for Multi-Antenna Wireless Systems" by Jens Jelitto and Gerhard Fettweis, introduces and evaluates a new space-time processing concept that reduces the signal dimension space at the receiver by exploiting spatial correlation properties of received signals. Since the antenna signals are treated as independent signal components when they arrive at the receiver, the proposed technique involves no special assumptions concerning the structure of the antenna. Furthermore, no feedback information is required at the transmitter. Although the new concept is studied in the context of single transmit and multiple receive antennas, it does not prevent the application of multiple transmit antennas or multiple data streams commonly applied in space-time coding schemes.

For radio LANs, sharing of bandwidth is essential because radio spectrum is not only expensive but also inherently finite. This is in contrast to wired networks, where bandwidth can be increased arbitrarily by adding extra cables. However, the broadcast nature of the wireless link poses a difficult problem for multiple access in that the success of a transmission is no longer independent of other transmission. To make a transmission successful, interference must be avoided or at least controlled. Otherwise, multiple transmissions may lead to collisions and corrupted signals. A multiple access (or medium access control) protocol is required to resolve these access contentions among nodes and transform a broadcast wireless LAN into a logical point-to-point network. In "Adequacy between Multimedia Application Requirements and Wireless Protocols Features," the authors describe and compare the medium access control mechanisms in the IEEE 802.11, Bluetooth, HomeRF, and HIPERLAN standards. In order to support multimedia applications, the authors suggest service differentiation and traffic classes to be provided by the underlying wireless network. To achieve these goals,

a priority-based multiple access protocol and the efficient management of different traffic queues are required.

The use of the Internet over mobile devices is becoming increasingly pervasive. However, a well-known problem for TCP connections over wireless links is that errors introduced by the wireless channel interfere with the TCP protocol, which operates well on a reliable physical link. Several methods to optimize TCP transmission over 802.11b wireless LANs are discussed in the article "Vertical Optimization of Data Terminals for Mobile Wireless Terminals," including improvements in power consumption, throughput, delay, and jitter. The performance improvements require modifications only on the mobile device, thereby allowing improved performance even when a mobile terminal roams into different wireless LANs.

The free-space wireless link is more susceptible to eavesdropping, fraud, and unauthorized transmission than its wired counterpart. Unauthorized people can tap the radio signal from anywhere within range. If someone sets a mobile terminal within a wireless coverage area to transmit packets endlessly, all other clients are prevented from transmitting, thus bringing the network down.

Being an open medium with no precise bounds makes it impractical to apply physical security like in wired networks. Nevertheless, several security mechanisms can be used to prevent unauthorized access of data transmitted over a wireless LAN. The article "Your 802.11 Wireless Network has No Clothes" by William Arbaugh *et al.*, discusses the flaws in the security mechanisms used by most access points supporting the IEEE 802.11 wireless standard. A key management system based on the Dynamic Host Configuration Protocol (DHCP) for mobile terminals as well as a higher-layer authentication system and the use of a higher-layer transport mechanism (e.g., IPSec) are proposed and evaluated. The combination of these mechanisms provides a robust interim solution until hardware supporting the new standards is deployed.

Being layer 2 technologies, wireless LANs offer limited Internet roaming capabilities and global user management features, including billing and identification features. The final article, "Cellular Access Control and Charging for Mobile Operator Wireless Local Area Networks" by Henry Haverinen, Jouni Mikkonen, and Timo Takamäki, describes a system that efficiently combines wireless LAN access with the widely deployed Global System for Mobile Communications/General Packet Radio Service (GSM/GPRS) roaming infrastructure. In addition, the architecture exploits GSM authentication, SIM-based user management, and secured billing mechanisms. This gives the cellular operator a major competitive advantage over Internet service providers who have neither a large mobile customer base nor a cellular-type roaming service. While service management, user trust, and global roaming are key factors in justifying the architecture presented, we also note the critical role assumed by the Extensible Authentication Protocol (EAP) in ensuring network security. The proposed architecture is currently being commercialized and is generic enough to be used on any access network that supports EAP.

The guest editors are grateful to the authors and reviewers for their hard work in ensuring that some very fine articles are published in this feature topic. It is hoped that the articles will stimulate more innovative research in this important subject area.

BIOGRAPHIES

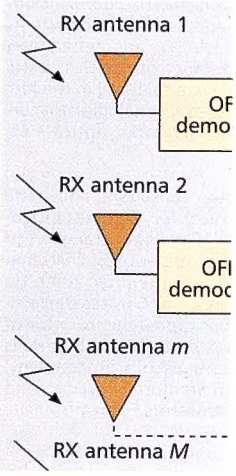
BENNY BING (bennybing@ieee.org) is a research faculty member with the School of Electrical and Computer Engineering at the Georgia Institute of Technology. He has published over 30 technical papers and six books, including *Wireless Local Area Networks*, which has been adopted by Cisco Systems worldwide. He is a technical consultant to several wireless and networking companies, and organizes the International Conference on Wireless LANs and Home Networks (www.icwlhn.org). Recently, he was featured in the *MIT Technology Review* in a special issue on wired and wireless technologies. His current research interests include wireless LANs, cable networks, optical networks, protocol design, and queuing theory.

CHRIS HEEGARD [F] is currently an independent investor, consultant, and cattle rancher. Previously he was the chief technology officer for the Texas Instruments Wireless and Home Networking Business Units. He co-founded and served as CEO of Alantro Communications, a company specializing in wireless LAN semiconductor technology, which was acquired by TI in September 2000. He served as a faculty member at the School of Engineering at Corness University for 19 years. He is the author of numerous publications, the inventor of several patents, and is a co-author of the first book on turbo coding. He is also founder of Native Intelligence, a digital communications software company. He received his electrical engineering degrees from Stanford University and the University of Massachusetts. He was a Texas Instruments Fellow.

BOB HEILE is a 20-year veteran in the field of data communications and wireless data with several articles and workshops to his credit. He is a founding member and chair of 802.15, the IEEE working group on wireless personal area networks, and is also one of the organizers for the 5 GHz global harmonization effort. In 1990 he was one of the founding members of 802.11. He is currently doing wireless communications consulting for several high profile companies. Before that he was with GTE responsible for wireless opportunity business development. He joined BBN in early 1997, prior to its acquisition by GTE, with the mission of commercializing wireless ad hoc networking and wireless PAN technologies. From 1990 to 1996 he was VP of engineering and business development for TyLink Corp., a bootstrap startup in high-speed digital access products and network and circuit management software, and was a co-founder of Windata, Inc., a developer and manufacturer of wireless LANs. From 1980 to 1990 he was with Codex, a subsidiary of Motorola, where he was VP/GM of the company's modem business. He holds a B.A. degree from Oberlin College, and M.A. and Ph.D. degrees in physics from The John Hopkins University.

CO-CHANNEL INTERFERENCE CANCELLATION BASED ON MIMO OFDM SYSTEMS

LUCA GIANGASPERO AND LUIGI AGAROSSI, PHILIPS RESEARCH MONZA
GIOVANNI PALTENGHI, CEFRIEL
SHUTAI OKAMURA, OSAKA UNIVERSITY
MINORU OKADA, INSTITUTE OF SCIENCE AND TECHNOLOGY
SHOZO KOMAKI, OSAKA UNIVERSITY



MIMO systems are currently stimulating considerable interest across the wireless industry as they appear to be a key technology for future wireless generations.

ABSTRACT

This article deals with the exploitation of multiple input multiple output (MIMO) systems for broadband wireless indoor applications. Two systems, the Wind-Flex and Ubiquitous Antenna, are considered. Both aim to increase the system capacity by different approaches, to increase the single link data rate and the number of users in the whole system, respectively. Computer simulation results show the effectiveness of both MIMO systems.

INTRODUCTION

The development of wireless communication systems for high-bit-rate data transmission and high-quality information exchange between terminals is becoming one of the new challenging targets in telecommunications research. The market demand for broadband multimedia services, ubiquitous networking, and Internet access via portable devices is expected to grow enormously, pushing the development of modem and system architectures for high-bit-rate transmission. Multiple input multiple output (MIMO) systems are currently stimulating considerable interest across the wireless industry because they appear to be a key technology for future wireless generations.

An (N,M)-MIMO wireless system can be generally defined as a MIMO system in which N signals are transmitted by N antennas at the same time using the same bandwidth and, thanks to effective processing at the receiver side based on the M received signals by M different antennas, is able to distinguish the different transmitted signals. The processing at the receiver is essentially efficient co-channel interference cancellation on the basis of the collected multiple information.

This permits improving system capacity whether the interest is to increase the single link data rate or increase the number of users in the whole system. The Wind-Flex MIMO and the Ubiquitous Antenna System are examples of these different ways to exploit MIMO properties.

For the Wind-Flex MIMO system [1], the target is to increase the capacity of a single link. The single user transmitter (Tx) and receiver (Rx) are equipped with N and M antennas, respectively. The basic idea is to usefully exploit the multipath rather than mitigate it, considering the multipath itself as a source of diversity that allows the parallel transmission of N independent substreams from the same user. This is quite different from a time-division multiplexing (TDM), frequency-eivision multiplexing (FDM), or code-division multiplexing (CDM) technique: there is no explicit orthogonalization of the N substreams. Instead, it is the rich multipath environment that makes the N substreams independent of each other. Since the user data is sent in parallel over N antennas, the effective bit rate is increased by a factor N.

To separate the N simultaneously transmitted signals at the receiver, the space time coding and space-division multiplexing techniques are used. Space time coding introduces a spatiotemporal correlation among transmitted signals to improve the information protection, while the goal of space-division multiplexing is to increase the data rate. Several different space time coding techniques have been considered: space time trellis codes [2], space time block codes [3], and D-BLAST [4]. All of them appear to be too complex for Wind-Flex-specific application, so space-division multiplexing [5] has been chosen. Among space-division multiplexing techniques, V-BLAST [6, 7] seems to exhibit the best trade-off between performance and complexity. The V-BLAST algorithm [6, 7] implements a nonlinear detection technique based on a spatial nulling process combined with symbol cancellation to improve performance. The spatial nulling process, based on the zero forcing (ZF) approach (a minimum mean square error approach is also possible), is used to separate the individual substreams. Conceptually, each substream in turn is considered to be the desired

1536-1284/02/$17.00 © 2002 IEEE

signal, and the remainders are considered as interferers to be cancelled (nulling). To improve performance, symbol cancellation is used in conjunction with spatial nulling. Symbol cancellation means that after the strongest substream has been detected, the contribution of this substream is subtracted from the received signals. Thus, the remaining weaker substreams are easier to recover since the strongest substream has been removed as a source of interference. This process is reiterated until all the substreams have been detected.

Different from Wind-Flex, the target of the Ubiquitous Antenna System [8, 9] is to increase system capacity as indicated by the total number of users in the system. The Ubiquitous Antenna System is composed of multiple microcellular radio base stations (RBSs) deployed over the service area and radio-on-fiber (RoF) link [10] which connects RBSs to the central control station (CCS). In this system, the transmitted signals from mobile terminals (MTs) are propagated through multipath channels and received at the plural of the RBSs. The received radio frequency (RF) signal at each RBS is modulated into the intensity of the optical carrier and the modulated optical signal is sent to the CCS via RoF link. At the CCS, all the optical signals from the RBSs are converted to RF signals again. Then the CCS performs all the demodulation sequences and signal processing. Since all the received signals at the RBSs are collected to the CCS, we can employ more sophisticated signal processing such as co-channel interference cancellation [11] and joint detection at the CCS. Hence, the Uubiquitous Antenna System can achieve space-division multiple access (SDMA), which allows multiple MTs to operate in the same time slot and frequency channel, and then improve the total number of users and frequency utilization efficiency of broadband wireless access systems.

Both systems adopt the orthogonal FDM (OFDM) modulation scheme even if they have different features concerning the number of subcarriers, subcarrier modulation, and so on. The choice of the OFDM is basically due to its interesting properties as a wideband modulation scheme and is generally accepted in the context of wireless indoor communications (e.g., IEEE 802.11a).

The article is organized as follows. In the next section the Wind-Flex system model and simulation results are presented. The third section is dedicated to the Ubiquitous Antenna System description and performance results. Concluding remarks are given in the final section.

Wind-Flex System

Wind-Flex Overview

The aim of the Wind-Flex project [12] is to design and develop a wireless high-bit-rate flexible and configurable modem architecture, which works in single-hop ad hoc networks and provides wireless access to the Internet for slowly moving users (about 1 m/s) in an indoor environment. A basic requirement for the Wind-Flex modem is the capability to manage variable bit rates, ranging adaptively from 64 kb/s to more than 100 Mb/s of payload, to bring new services and trigger new interesting and appealing applications with different quality of service (QoS) requirements. The 17 GHz unlicensed frequency band has been chosen since the frequency bands around 2.4 and 5 GHz have already been allocated and are close to saturation. The available band of 200 MHz (17.1–17.3 GHz) is divided into four 50 MHz wide channels not simultaneously selectable. Due to the adopted carrier frequency, the coverage ranges from 5 m for non-line of sight (NLoS), to 20 m for line of sight (LoS) using omnidirectional antennas. Subcarrier modulation schemes are binary phase shift keying (BPSK), quaternary PSK (QPSK), 16-quadrature amplitude modulation (QAM), and 64-QAM. The constellation of each subcarrier is adaptively chosen among the various schemes according to the subcarrier signal-to-noise ratio (SNR), the target bit error rate (BER), and medium access control (MAC) requests. With respect to the channel coding, the coding scheme of the Wind-Flex modem is a parallel convolutional turbo code. The code rates established are 1/2, 2/3, and 3/4. At the MAC layer of the Wind-Flex system, the bit transmission is organized in a way time-division multiple access/time-division duplex (TDMA/TDD). The upper interface of DLC layer will serve Internet Protocol version 4 (IPv4) and version 6 (IPv6). This selection gives the widest possible range of potential applications to the Wind-Flex modem.

System Model

MIMO techniques are based on the assumption of a flat fading channel. This requirement is obviously not verified in a 50 MHz wide wireless indoor channel, such as the Wind-Flex one. However, the use of OFDM modulation makes the flat fading hypothesis true for each OFDM subband, allowing exploitation of the MIMO approach for broadband wireless applications as well.

The transmitter is equipped with N antennas (Fig. 1). A traditional channel encoder encodes the source information bits. The coded bits are then mapped on the symbols of the constellation adopted for each OFDM subcarrier. The vectorial nature of the transmission is introduced by demultiplexing ($1 \rightarrow N$) these symbols. In the considered architecture this demultiplexer represents the space encoder. It maps symbols on the *N space channels*, which are substreams of the same user. A serial to parallel converter for each space channel takes L of these symbols to form the input for the OFDM modulator. L is the number of *frequency channels*, that is, the number of OFDM subcarriers (in WIND-FLEX L = 128). To avoid any intersymbol interference (ISI) due to the delay spread of the channel, a cyclic prefix is appended to each OFDM symbol. The corresponding antenna transmits the output from each modulator.

This structure permits the simultaneous transmission of $N \cdot L$ M-QAM symbols in the same bandwidth and with the same total transmitted power required by an OFDM symbol to carry only L of them.

The receiver is equipped with M antennas (note that V-BLAST requires $M \geq N$). Each

The use of OFDM modulation makes the flat fading hypothesis true for each OFDM sub-band, allowing the exploitation of the MIMO approach also for broadband wireless applications.

We assume that the channel remains fixed during a frame, but it randomly changes from frame to frame. This is a totally reasonable assumption in wireless applications where there is limited mobility, such as the Wind-Flex one.

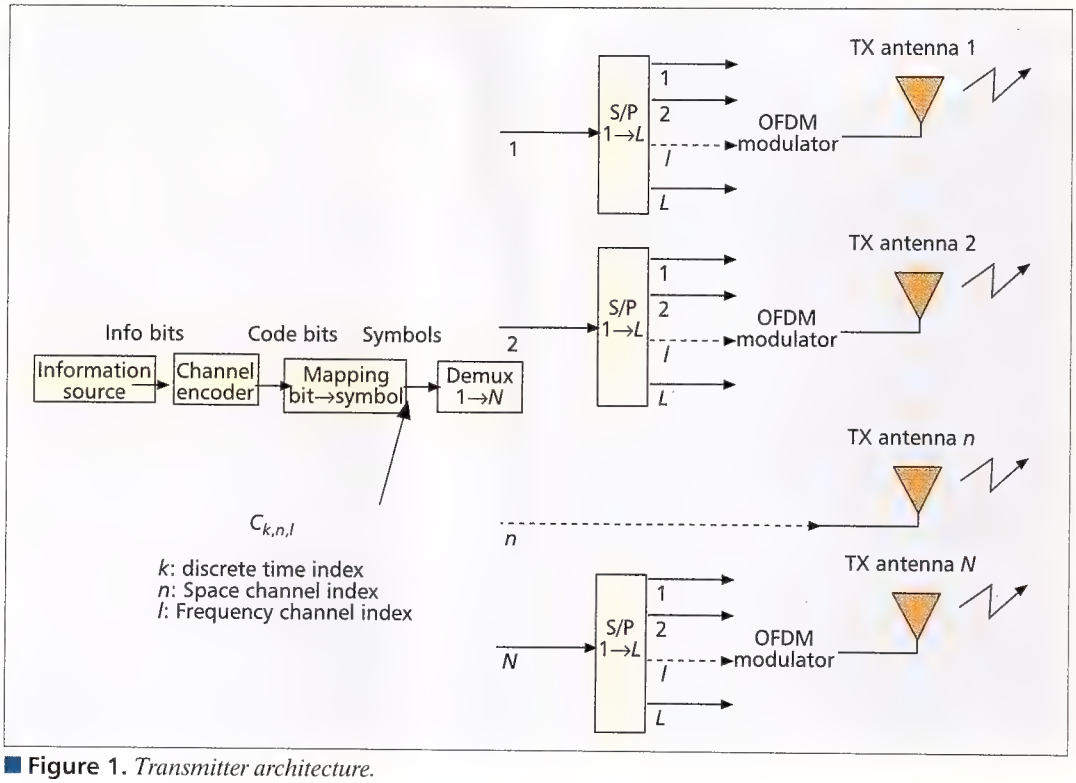


Figure 1. *Transmitter architecture.*

antenna receives a different noisy superposition of the faded versions of the N transmitted signals (Fig. 2a). If the antennas are sufficiently spatially separated (more than $\lambda/2$) and there is sufficiently rich scattering, the transmitted signals arriving at different receive antennas undergo independent fading. Moreover, if the channel response is known at the receiver, V-BLAST algorithm is able to detect the N transmitted signals. Channel response can be estimated at the receiver using a training sequence embedded in each TDMA frame. In so doing, we assume that the channel remains fixed during a frame, but it randomly changes from frame to frame. This is a totally reasonable assumption in wireless applications where there is limited mobility, such as Wind-Flex.

At each OFDM symbol time, the output of the OFDM demodulator corresponding to the receive antenna m is a set of L signals, $r_{m,l}$, one for each frequency channel

$$r_{m,l} = \sum_{n=1}^{N} H_{m,n,l} \cdot C_{n,l} + \eta_{m,l} \quad \text{with } l = 1, \ldots, L$$

where $H_{m,n,l}$ is the complex coefficient representing the frequency response of the channel from the transmit antenna n to the receive antenna m at multicarrier frequency l, and $\eta_{m,l}$ are independent samples of a complex Gaussian random variable with zero mean and variance N_0, representing noise (note that N_0 is the variance of the noise at the receiver input [6]).

The M outputs related to the multicarrier frequency l are the inputs to a V-BLAST signal processor that detects the N different M-QAM symbols transmitted in this frequency channel. There are L V-BLAST signal processors, each operating in a specific frequency channel (Fig. 2b). The $N \cdot L$ M-QAM symbols, obtained at the output of the L V-BLAST signal processors, are then serialized by a parallel to serial converter in order to complete the traditional receiver processing.

Concerning the channel model, the single input single output (SISO) model adopted to calculate the $H_{m,n,l}$ coefficients is the one presented in [13]. The NLoS channel is modeled with 18 taps. The taps' amplitudes have been shaped following different probability density functions (pdfs): a combination of exponential and Weibull pdfs for the first bin and exponential pdfs for the others. In the considered context, however, the SISO model is not sufficient to completely characterize the channel, since the MIMO approach also entails consideration of the spatial correlation characteristics of the transmission system. Several studies have recently demonstrated that if channel path gains of a (N,M) MIMO system are independent, the channel capacity scales linearly with n, where $n = \min(N,M)$ [14]. The problem is that in real propagation conditions, these channel coefficients could be partially correlated. Generally, the correlation of the coefficients depends on many factors such as the physical parameters of transmit and receive antennas (e.g., antenna spacing), and the characteristics and distribution of the scatterers. It has been observed that when paths are correlated, the channel capacity can be significantly smaller than when they are not. This suggests that further research is required into the spatial correlation problem [15].

In this work, the correlation has been taken into account by means of a synthetic parameter (i.e., the mean correlation coefficient of the MIMO channel). Some correlation matrices K (the generic element $k_{i,j}$ of the correlation matrix

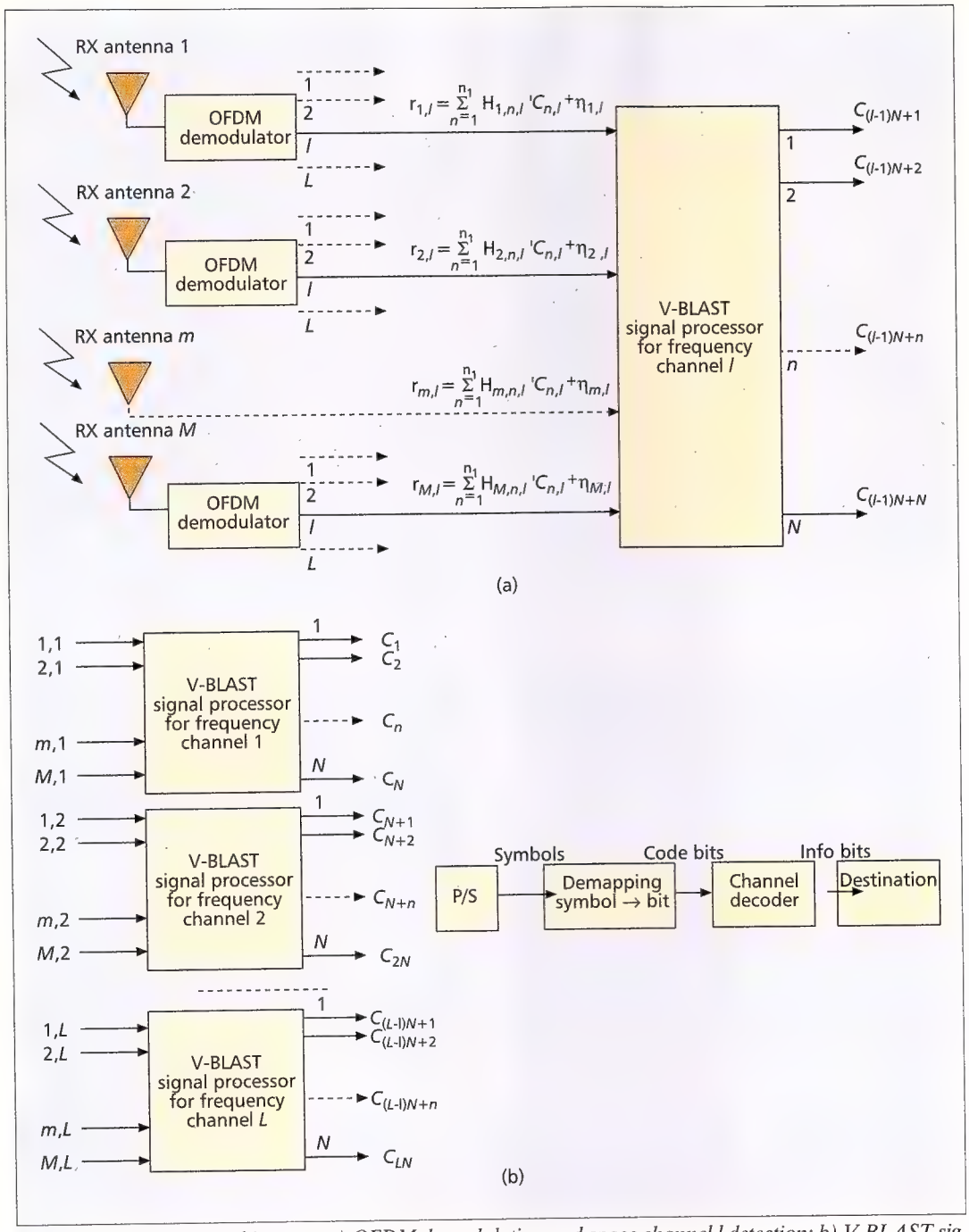


Figure 2. *Receiver architecture: a) OFDM demodulation and space channel l detection; b) V-BLAST signal processors detection and traditional receiver processing.*

is the correlation coefficient between the *i*th and *j*th path gains) have been defined with an increasing value of the mean correlation coefficient ranging from 0 (totally uncorrelated path gains) to 0.8 (almost completely correlated path gains) with step granularity of 0.2.

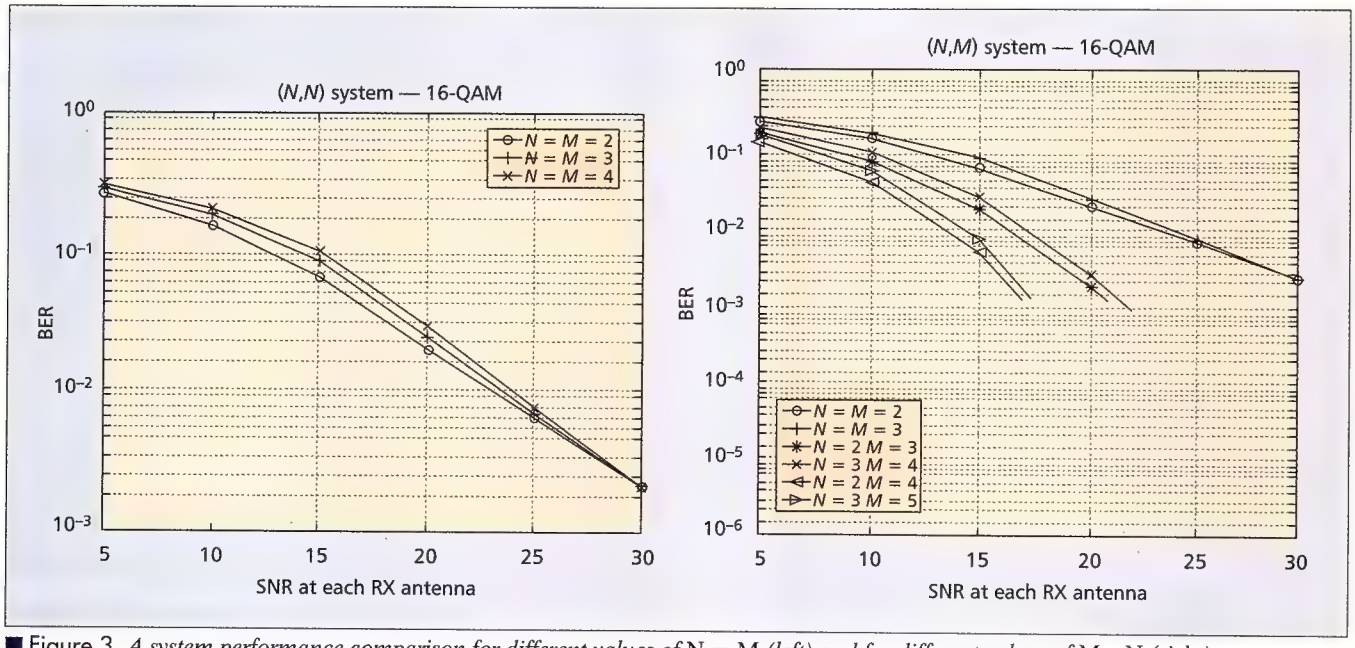
Simulation Results

In order to verify the performance of the considered architecture in ideal and nonideal propagation conditions (i.e., the presence of spatial correlation between path gains), a simulator, based on Wind-Flex specifications, has been implemented. In our simulations, we have assumed that the received signals are corrupted by additive white Gaussian noise. We have also assumed ideal symbol and sample-clock synchronization at the receiver. We have considered noncoded system performance. The BER, as a function of the SNR at each receive antenna, has been evaluated.

The first set of performance curves refers to ideal propagation conditions (i.e., the path gains can be considered independent). The impact on the performance of N and M has been analyzed.

First, a system having the same number of transmit and receive antennas has been simulated. In Fig. 3 (left) the performance of this system, with 16-QAM on each subcarrier, for different N is reported.

In our simulations, we have assumed that the received signals are corrupted by additive white Gaussian noise. We have also assumed ideal symbol and sample-clock synchronization at the receiver. We have considered the non-coded system performance.



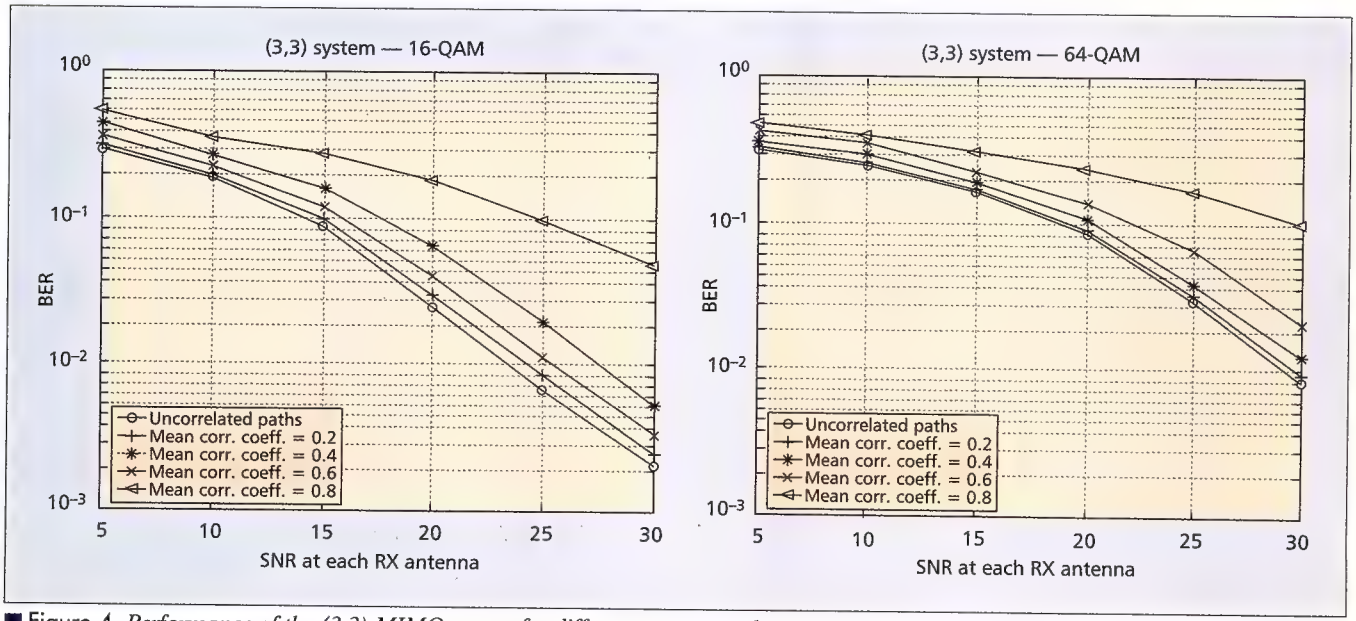
■ Figure 3. *A system performance comparison for different values of* N = M *(left) and for different values of* M – N *(right).*

This figure proves that the V-BLAST technique permits increasing the bit rate without significantly worsening the BER. Remember, in fact, that in ideal propagation conditions the bit rate of the SDM systems scales linearly with the number of transmit antennas. It is important to note that the total transmitted power and used bandwidth are constant.

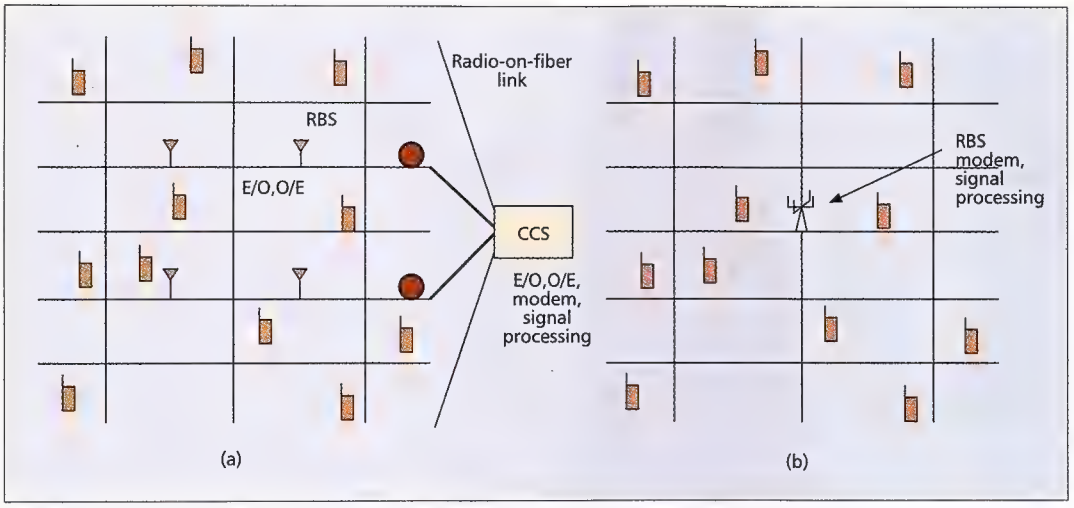
The performance curves of a MIMO system with different numbers of receive antennas, *M*, using 16-QAM on each subcarrier, have also been evaluated (Fig. 3, right). As expected, increasing *M* improves performance. This is due to the availability of a greater number of received signals, which can be combined in a more efficient way to obtain a more accurate estimate of the transmitted signals. Moreover, some other simulations indicate that the performance improvement is a function of the difference $M - N$, resulting in being substantially independent from *N* or *M* separately.

Concerning the nonideal propagation condition, the analysis has been limited to the (2,2) and (3,3) MIMO systems. Figure 4 shows the simulation results for a (3,3) system with different values of the mean correlation coefficient among paths. The mean correlation coefficient ranges from 0 to 0.8 with steps of 0.2.

The analysis of these curves shows an interesting nonlinear relation between performance degradation and the mean correlation coefficient. Moreover, the negative effect of the spatial correlation becomes more and more evident as SNR increases. Under a threshold SNR, BER performance is limited mainly by the presence of noise. The curves show that the SNR



■ Figure 4. *Performance of the (3,3) MIMO system for different mean correlation coefficients and modulations.*



■ **Figure 5.** *General description of (a) the ubiquitous antenna system and (b) the centralized antenna system.*

loss of the (3,3) system due to spatial correlation does not exceed 5 dB up to a mean correlation coefficient equal to 0.6 in the considered SNR range. It must be pointed out that 0.6 is a high value of correlation. Beyond this value the use of a MIMO system would be inefficient for the specific propagation environment. The results for the (2,2) system exhibit similar properties.

Ubiquitous Antenna System

Ubiquitous Antenna System Overview

A general description of the Ubiquitous Antenna System is shown in Fig. 5a. The system is composed of multiple microcellular RBSs deployed over the service area, a CCS, and the RoF link. Each RBS has only electrical-to-optical (E/O) and optical-to-electrical (O/E) converters, and requires no RF modulation or demodulation functions. The radio signals received at the RBSs are applied to the E/O converter to modulate the intensity of the optical carrier, and they are sent to the CCS via the RoF link, maintaining the radio signal waveform. The CCS converts the optical signals from RBSs to electrical signals again by corresponding O/E converters, and then performs all the signal processing and demodulation sequences. Since all the signals received at the RBSs are obtained at the CCS, co-channel interference cancellation and joint detection can be performed at the CCS similar to the adaptive array antenna systems [11]. The ubiquitous antenna system enables the multiple MTs to operate at the same frequency channel simultaneously by making effective use of joint detection. Thus, it does not require frequency allocation for each RBS. That is, all the RBSs can use a whole band assigned to the system. Hence, the ubiquitous antenna system achieves higher frequency utilization efficiency whatever the assigned bandwidth is strictly limited. Furthermore, since the CCS has expensive components such as RF components, modulator, demodulator, and signal processing components such as a joint detector, and RBSs requires O/E and E/O converters, the ubiquitous antenna system allows us to reduce the total cost to deploy and maintain the system.

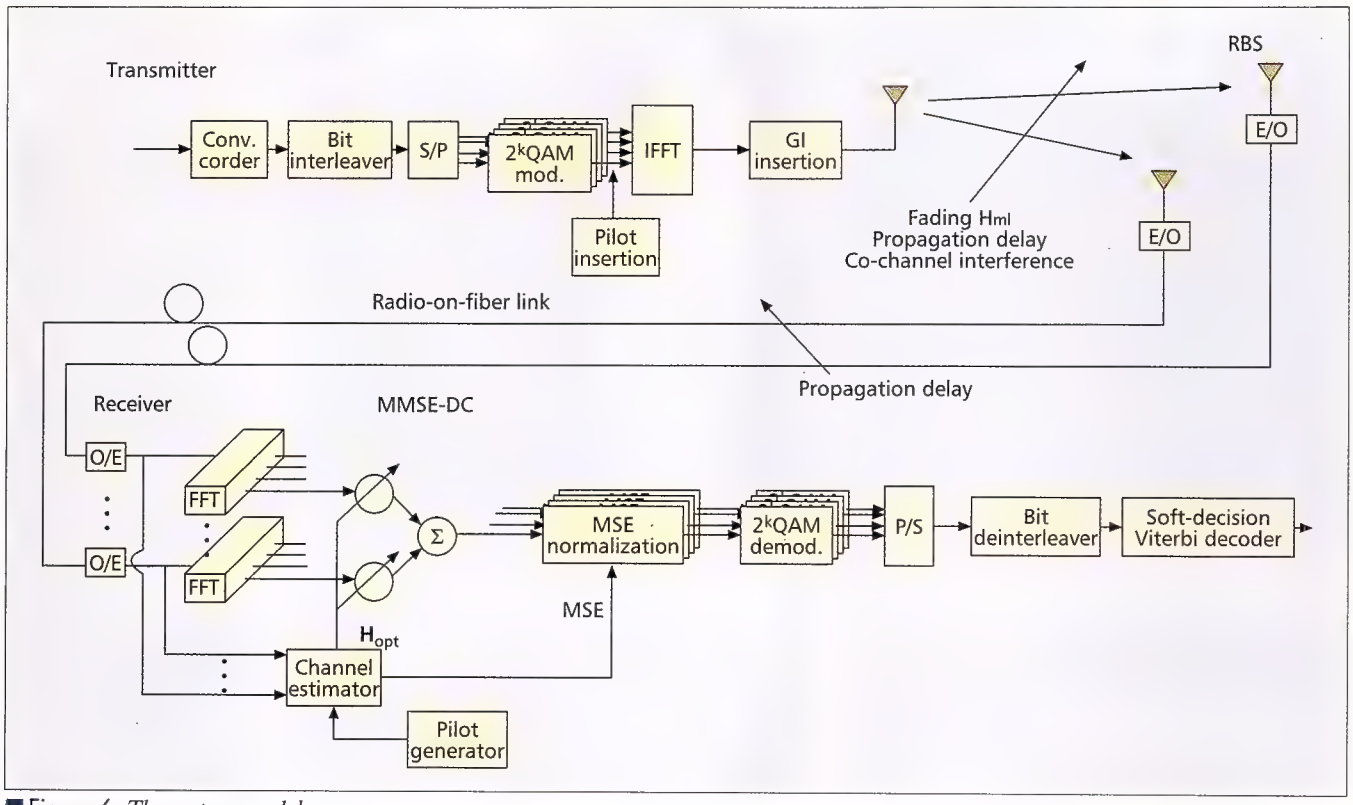
A ubiquitous antenna system is also known as a distributed antenna system, which has been studied in recent years [16, 17]. In a distributed antenna system, antenna elements are distributed over the service area as a microcellular RBS and their outputs are brought, in analog form, to a central server. Similar to the adaptive array, interference cancellation is performed at the central server. Since the distance between antenna elements is much longer than in a conventional centralized adaptive array antenna system (Fig. 5b), we can further obtain macrodiversity. In [10], Clark *et al.* evaluated and compared centralized and distributed antenna systems. From their remarks, the distributed antenna system allows us to obtain improvement in system capacity.

System Model

In order to evaluate the performance of the ubiquitous antenna system, a system model is considered in this article. The system model is illustrated in Fig. 6. In the following, we concentrate on the uplink connection. Let us assume that N MTs in the service area transmit COFDM signals simultaneously at the same frequency channel. At the nth MT, the binary streams are error correction encoded and interleaved over each COFDM subcarrier. The encoded binary streams are mapped onto the QAM symbols at the modulator. In order to estimate the channel impulse response (CIR) between each MT and each RBS, pilot symbols are attached before data symbols. The symbols are then inverse discrete Fourier transformed (IDFT). The guard interval, also known as the cyclic extension, is inserted in order to remove ISI due to delay spread. After that, the signal is transmitted to the RBSs through the radio propagation channel.

In a radio propagation path, the transmitted signal from each MT is affected by fading, co-channel interference, path loss, and propagation delay. The signals are received at the M RBSs deployed over the service area. The received signal at each RBS is corrupted by additive white

The ubiquitous antenna system enables the multiple MTs to operate at the same frequency channel simultaneously by making effective use of joint detection. Thereby, it does not require the frequency allocation for each RBS. That is, all the RBSs can use a whole band assigned to the system.



■ Figure 6. *The system model.*

Gaussian noise (AWGN) as well as co-channel interference from other MTs. The received signal is converted to an optical signal by an E/O converter, and sent to the CCS via the RoF link. Since the spacing among RBSs is hundreds of meters, about several hundred nanoseconds of delay difference arises in the RoF link. In general, this delay difference could be a problem in wideband transmission. Fortunately, since the COFDM signal has a guard interval inserted at the head of each COFDM symbol in order to avoid ISI and interchannel interference due to delay spread, the delay difference can be ignored.

At the CCS, the O/E converters convert the optical signals from RBSs to electrical signals again.

For the sake of notational convenience, we define the M-dimensional received signals vector as $\mathbf{y}^k = [\mathbf{y}_1^k, \mathbf{y}_2^k, \ldots, \mathbf{y}_M^k]^T$, where k denotes the subcarrier index. The received vector is then given by

$$\mathbf{y}^k = \mathbf{H}^k\mathbf{x}^k + \mathbf{z}^k,$$

where $\mathbf{x}^k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_N^k]^T$ is the N-dimensional transmitted signal vector, $\mathbf{z}^k = [\mathbf{z}_1^k, \mathbf{z}_2^k, \ldots, \mathbf{z}_M^k]T$ is the N-dimensional AWGN vector, where the elements $\mathbf{z}_m^k$ are independent and identically distributed Gaussian random variables with zero mean and variance σ_n^2. $\mathbf{x}^T$ denotes the transpose of $\mathbf{x}$. The frequency response matrix $\mathbf{H}^k = [\mathbf{h}_{mn}^k]$ is an $M \times N$ matrix whose element $\mathbf{h}_{mn}^k$ denotes the response between the nth MT and the mth RBS. In this article, we assume that the frequency responses for different MTs and RBSs are statistically independent, stationary, and complex Gaussian random variables.

Joint Detection — As a joint detector, we employ the minimum mean square error diversity combiner (MMSE-DC). The MMSE-DC performs the joint detection by using M signals form RBSs. Before the MMSE-DC, the received signal from each RBS is divided into each subcarrier by the corresponding FFT processor. Then the MMSE-DC is performed subcarrier by subcarrier. The received signal from each RBS is weighted by the optimum weight matrix given by

$$\mathbf{H}_{\text{opt}}^k = (\mathbf{H}^k)\mathbf{H}(\mathbf{R}_{yy}^k)^{-1},$$

where $\mathbf{R}_{yy}^k$ is $M \times M$ correlation matrix of received signals. The weighted signals are then combined. The output of the MMSE-DC is given by

$$\mathbf{x}^k = \mathbf{H}_{\text{opt}}^k y^k,$$

where $\mathbf{x}$ is the estimated transmitted signals vector.

Channel Estimation — Not only can the MMSE-DC-based joint detection described previously mitigate the performance degradation due to fading, it can also remove the co-channel interference transmitted from other MTs. However, in order to establish joint detection, estimation of CIR is required at the receiver. In the proposed system, each MT inserts the pilot symbol satisfying the optimum MSE condition discussed in [18]. The receiver estimates the CIR by calculating the correlation between the received signal at each RBS and reference pilot signal and the estimated CIR. Then the receiver estimates $\mathbf{H}_n^k = [\mathbf{h}_{1n}^k, \mathbf{h}_{2n}^k, \ldots, \mathbf{h}_{mn}^k]^T$ by transforming CIR into the frequency domain using FFT, where $\mathbf{H}_n^k$ is an M-dimensional channel response vector.

MSE Normalization — At the output of the MMSE-DC, all subcarriers have the same desired signal power, while the noise power at each subcarrier is different. On the other hand, the following Viterbi decoder uses the Euclidean distance as the path metrics. That is, the decoder is optimum in terms of minimizing the BER when all the noises are independent and identically distributed (i.i.d.) Gaussian random variables. In order to perform the Viterbi decoder in the optimum condition, the output signal of the MMSE-DC is normalized by the MSE or the noise variance of the corresponding signal. The MSE for the nth MT is given by

$$\mathrm{MSE}_n^k = \sigma_d^2 - (\mathbf{H}_n^k)^{\mathbf{H}}(\mathbf{R}_{yy}^k)^{-1}\mathbf{H}_n^k,$$

where σ_d is the desired signal power. This normalization is performed subcarrier by subcarrier as well as for the MMSE-DC. The noise at the output of the MSE normalization is i.i.d., and the Viterbi decoder performs as the optimum decoder.

Then the normalized signals are demodulated and deinterleaved. The deinterleaved signal is then applied to the soft-decision Viterbi decoder. Finally, we can obtain the desired user's binary streams.

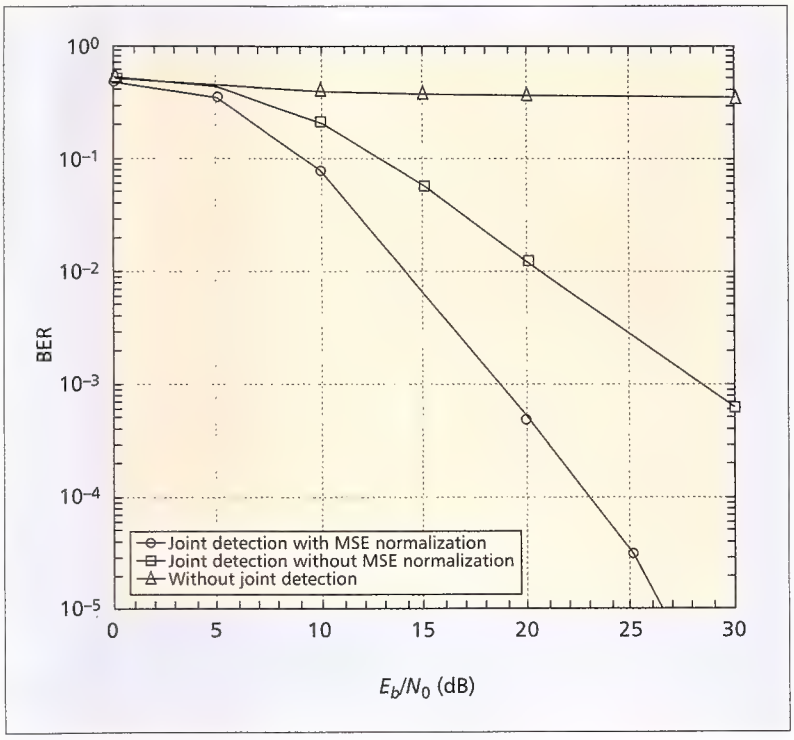
Simulation Results

In this section we analyze the performance of the proposed ubiquitous-antenna-based wireless LAN system by computer simulations. In the following, we evaluate the BER and frequency utilization efficiency of uplink connection in comparison with other systems.

The COFDM parameters are based on the IEEE 802.11a standard. The entire channel bandwidth is divided into 64 subchannels. The 48 subchannels are used to transmit data. Differential quadrature phase shift keying (DQPSK) is employed as a subchannel symbol modulation format. The symbol duration is 4.2 μs including 1.0 μs of guard interval, which is just a little bit longer than that of the WLAN standards in order to mitigate performance degradation due to the delay difference among RoF links. For forward error correction (FEC), half-rate convolutional error correction coding with constraint length of 7 (64-state) is employed. In the system, each MT can transmit a signal with a data rate of 11.5 Mb/s over a 15 MHz channel, and its transmission efficiency is approximately 0.77 b/s/Hz.

As a propagation channel, two-sample spaced two-ray Rayleigh fading channel with Doppler shift, f_d = 40 Hz, is assumed. The interval between two rays is 150 ns. The path loss exponent is 4.0. We also assume the symbol timing of the COFDM signals is synchronized at every MT. We ignore the frequency offset between the MTs and CCS local oscillators, and nonliner distortion due to the RoF link.

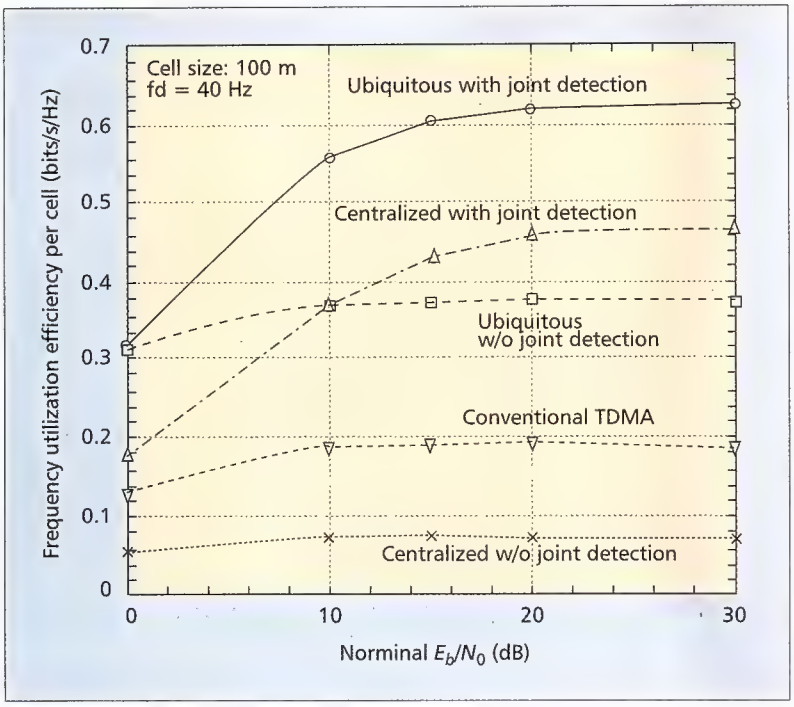
First, we analyze the BER performance of the ubiquitous antenna system in order to demonstrate the basic characteristics of the proposed system and the effect of the MSE normalization scheme. In this simulation, we assume that two RBSs are deployed in the horizontal axis, and each RBS is connected to the CCS by the RoF link. The distance between RBSs is 100 m, and it causes 500 ns of delay difference due to the RoF link. Two MTs are located at the center of two RBSs. The BER performance of the proposed system against E_b/N_0 is shown in Fig. 7. The BER performance without joint detection is intolerable for transmitting data. However, the ubiquitous antenna system with joint detection can drastically improve the BER performance. Furthermore, in contrast to the performance without joint detection, the proposed system without MSE normalization drastically improves BER performance. Furthermore, MSE normalization gives a gain of 10 dB at BER = 10^{-3} by making effective use of the implicit subcarrier diversity effect.



■ Figure 7. *BER performance of the ubiquitous antenna system.*

Next, we analyze the frequency utilization efficiency of the ubiquitous antenna system. The allocation of the RBSs is shown in Fig. 5a. In this simulation, we assume 16 cells, and the ubiquitous antenna system is composed of four RBSs corresponding to the four cells in the center of the area. Each RBS is located at the center of the cell and connected to the CCS with the RoF link. Each cell size is 100 m. All the MTs transmit COFDM signal simultaneously at the same frequency channel. The ubiquitous antenna system only demodulates the signals from MTs in the central four cells, and signals from the other MTs around the four cells are considered co-channel interference. The ubiquitous antenna system performs the MMSE-DC-based joint detection by using the signals received at four RBSs.

For comparison purposes, we evaluate the performance of the ubiquitous antenna system without joint detection. In this case, all the RBSs have their own demodulators and each RBS demodulates the received signals independent of the other RBSs. All the MTs are still



■ Figure 8. *Frequency utilization efficiency.*

operating at the same frequency. Furthermore, we evaluate the macrocell system, or one RBS located at the center of the service area as shown in Fig. 5b. In the following, we call it the central antenna system. We assume the central antenna system with and without joint detection. With joint detection, the RBS has a four-element array antenna, and MMSE-DC-based joint detection is performed using the array antenna. In the central antenna system without joint detection, however, the RBS has only one antenna element, and no joint detection is performed. As in the ubiquitous antenna system, all the MTs operate at the same frequency simultaneously. Moreover, we also evaluate the conventional digital communication system in which all the MTs access the RBS, located at the center of the four cells and equipped with only one antenna element, in TDMA mode. In all cases, all the MTs are uniformly distributed over the service area. In this simulation, we assume the transmission is successful when there is no bit error in a packet composed of a pilot symbol and 10 information symbols.

Figure 8 shows the frequency utilization efficiency against nominal E_b/N_0. The transmission power of an MT is determined so that E_b/N_0 per branch equals nominal E_b/N_0 at one RBS antenna that is 70 m apart from the corresponding MT. In this simulation, we assume four MTs are in the four cells. Furthermore, eight MTs are on the outside of these four cells as co-channel intercell interferers. In the conventional system, the efficiency is about 0.2 b/s/Hz since it allows only one MT to operate at the same frequency channel at one time. In the centralized system, it improves the efficiency to 0.5 b/s/Hz by using joint detection. On the other hand, in the ubiquitous antenna system, the efficiency is improved to 0.65 b/s/Hz by making effective use of joint detection. Furthermore, even the ubiquitous antenna system with selection diversity can achieve higher efficiency than a centralized one with joint detection in the lower E_b/N_0 region.

Conclusions

This article focuses on two MIMO OFDM-based systems, Wind-Flex and the ubiquitous antenna system. Their capability to increase overall system capacity has been investigated and the simulation results on their performance are reported.

The Wind-Flex system, based on the V-BLAST MIMO and OFDM modulation scheme, proves capable of greatly improving bit rate without increasing total transmitted power or required bandwidth. The simulation results confirm this feature for both ideal and nonideal propagation conditions.

The ubiquitous antenna system, which is composed of multiple RBSs deployed over the service area, the CCS, and an RoF link, allows all the MTs to operate at the same frequency simultaneously. The computer simulation results show the ubiquitous antenna system is capable of improving overall system capacity. This property is very important in designing wireless networks, especially when the radio resource is strictly limited.

References

[1] L. Giangaspero, G. Paltenghi and L. Agarossi, "A MIMO architecture for Wireless Indoor Applications," *Proc. IEEE Int'l. Conf. Wireless LANs and Home Nets.*, Singapore, Dec. 2001, pp. 317–26.

[2] V. Tarokh, N. Sehadri, and A. R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communications: Performance Criterion and Code Construction," *IEEE Trans. Info. Theory*, vol. 44, Mar. 1998, pp. 744–65.

[3] S. M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," *IEEE JSAC*, vol. 16, no. 8, Oct. 1998, pp. 1451–58.

[4] G. J. Foschini, "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-Element Antennas," *Bell Labs Tech. J.*, Autumn 1996.

[5] A. van Zelst, "Space Division Multiplexing Algorithms," *10th Mediterranean Electrotech. Conf.*, vol. 3, pp. 1218–21.

[6] P. W. Wolniansky *et al.*, "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel," invited paper, *Proc. ISSSE '98*, Pisa, Italy, Sept. 1998.

[7] G. J. Foschini *et al.*, "Simplified Processing for High Spectral Efficiency Wireless Communication Employing Multi-Element Arrays," *IEEE JSAC*, vol. 17, no. 11, Nov. 1999, pp. 1841–52.

[8] M. Toyama, M. Okada, and S. Komaki, "Maximal Ratio Combining Macro Diversity for Micro-Cellular Slotted ALOHA," *IEICE B-I*, vol. J79-B-I, 5, May 1996, pp. 271–77.

[9] S. Okamura, M. Okada, and S. Komaki, "On the Performance of the Ubiquitous Antennas for the Reception of COFDM Signals," *Proc. IEEE Int'l. Conf. WL LANs and Home Nets.*, Singapore, Dec. 2001, pp. 295–304.

[10] S. Komaki *et al.*, "Proposal of Radio Highway Networks for Future Multimedia-Personal Wireless Communications," *ICPWC '94*, Bangalore, India, Aug. 1994, pp. 204–8.

[11] Y. Li and R. Sollenberger, "Adaptive Antenna Arrays for OFDM Systems With Cochannel Interference," *IEEE Trans. Commun.*, vol. 47, no. 2, Feb.1999, pp. 217–29.

[12] http://www.vtt.fi/ele/research/els/projects/windflex.htm

[13] M. Lobeira *et al.*, "Parameter Estimation and Indoor Channel Modelling at 17 GHz for OFDM-Based Broadband WLAN," *IST Mobile Commun. Summit 2000*, Galway, Ireland, Oct. 2000, pp. 29–33.

[14] G. J. Foschini and M. J. Gans, "On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas," *WL Pers. Commun.*, vol. 6, no.3, March 1998, pp. 311–35.

[15] D.-S. Shiu *et al.*, "Fading Correlation and Its Effects on the Capacity of Multielement Antenna Systems," *IEEE Trans. Commun.*, vol. 48, no. 3, Mar. 2000, pp. 502–13.
[16] M. V. Clark *et al.*, "Distributed Versus Centralized Arrays in Broadband Wireless Networks," *Proc IEEE. VTC*, Rhodes, Greece, MA1-2, May 2001.
[17] A. Sklavos *et al.*, "Joint Channel Estimation in Multi-User OFDM System," *Proc. 6th Int'l. OFDM Wks.*, Sept. 2001.
[18] Y. Li, N. Seshadri, and S. Ariyavisitakul, "Channel Estimation for OFDM Systems with Transmitter Diversity in Mobile Wireless Channels," *IEEE JSAC*, vol. 17, Mar. 1999, pp. 461–71.

Biographies

LUCA GIANGASPERO (luca.giangaspero@philips.com) graduated in electronic engineering, telecommunications specialization, from the Polytechnic of Bari in 2001. In the same year he received a Master's degree in information technology, in the field of advanced transmission systems, from CEFRIEL, Polytechnic of Milan, where he worked on MIMO wireless transmission systems. He is now a research scientist at Philips Research Monza in the wireless systems and terminals area. He is currently involved in an IST European research project aimed at defining a flexible high-rate wireless modem for indoor applications. His main research interests are adaptivity and reconfigurability for future-generation wireless systems.

LUIGI AGAROSSI (luigi.agarossi@philips.com) received a degree in electronic engineering from the University of Bologna in 1981. He started working as a research scientist in the GTE Telecommunications Radio Link R&D Department, within the area of baseband and IF signal processing. In 1987 he joined Philips Research Central Laboratories (NATLAB), Eindhoven, The Netherlands. After one year he moved to Philips Research Monza Laboratory, Italy, where he started working on storage systems and architecture. He has experience in optical storage, optical channel modeling, and detection for high-density optical recording. He is the author of international patents and papers, and has taken part in many European projects. Since 1997 he has been working on wireless systems and terminals, and is currently involved with the Wireless Indoor Flexible High Bitrate Modem Architectures (Wind-Flex) European project as a project leader. Current interests are digital baseband processing and modulation for high-capacity systems, software-defined radio (SDR), and adaptive and reconfigurable radio.

GIOVANNI PALTENGHI (palteng@cefriel.it) received a degree in electronic engineering (summa cum laude) from the University of Brescia in 1999. He has been with CEFRIEL-Politecnico di Milano since 1999 and currently is a researcher in the Advanced Mobile and Wireline Transmission Systems unit. His main research interests are in the fields of residential/SOHO wireless systems and fiber optic communication systems. In particular, he is carrying out research on OFDM modulation, MIMO transmission systems, optical transport networks, and optical devices for "intelligent" WDM networks (optical add/drop multiplexers and optical crossconnects). He is a member of WIND-FLEX, an IST European project devoted to the design of a high-bit-rate flexible and configurable wireless architecture for indoor environments.

SHUTAI OKAMURA (okamura@roms.comm.eng.osaka-u.ac.jp) received a B.E. degree in electrical and electronic engineering from Shizuoka University, Japan, in 2000, and an M.E. degree from Osaka University, Japan, in 2001. He is currently pursuing a Ph.D. degree at Osaka University and engaged in research on radio communication systems. He is a member of the Institute of Electronics and Information Communication Engineers (IEICE) of Japan.

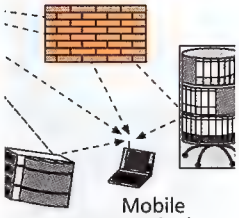
MINORU OKADA (mokada@is.aist-nara.ac.jp) received a B.E. degree in communications engineering from the University of Electro-Communications, Tokyo, Japan, in 1990, and M.E. and Ph.D. degrees, both in communications engineering, from Osaka University, Japan, in 1992 and 1998, respectively. In 1998 he joined the Department of Communications Engineering, Osaka University, as a research associate. From 1999 to 2000, he was with the University of Southampton, United Kingdom, as a visiting research fellow. In 2000 he joined the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, where he is currently an associate professor. He is a member of the IEICE, IEEE, and Institute of Image Information and Television Engineers of Japan (ITE). He received the Young Engineer Award from IEICE in 1999.

SHOZO KOMAKI [SM] (komaki@comm.eng.osaka-u.ac.jp) received B.E., M.E., and Ph.D. degrees in electrical communication engineering from Osaka University, in 1970, 1972, and 1983, respectively. In 1972, he joined NTT Radio Communication Laboratories, where he was engaged in repeater development for a 20 GHz digital radio system, and 16-QAM and 256-QAM systems. From 1990 he moved to Osaka University, Faculty of Engineering, and engaging in the research on radio and optical communication systems. He is currently a professor at Osaka University. He is a member of the IEICE and (ITE). He was awarded the Paper Award and the Achievement Award of IEICE in 1977 and 1994, respectively.

The Wind-Flex system, based on the V-BLAST MIMO and OFDM modulation scheme, shows the capability to greatly improve the bit-rate without increasing the total transmitted power or the required bandwidth. The simulation results confirm this feature for both ideal and non-ideal propagation conditions.

REDUCED DIMENSION SPACE-TIME PROCESSING FOR MULTI-ANTENNA WIRELESS SYSTEMS

JENS JELITTO, IBM ZURICH RESEARCH LABORATORY
GERHARD FETTWEIS, DRESDEN UNIVERSITY OF TECHNOLOGY



The need for wireless communication systems has grown rapidly during the last few years. Moreover, there is a steady growth in the required data rates. To meet those requirements, current and next-generation wireless systems and networks such as wireless LANs will support much higher data rates compared with established standards.

ABSTRACT

The need for wireless communication systems has grown rapidly during the last few years. Moreover, there is a steady growth in the required data rates due to the fact that more and more users request high-bit-rate services. To meet those requirements, current and next-generation wireless systems and networks such as wireless LANs (e.g., IEEE 802.11a) will support much higher data rates compared with established standards. This is basically done by applying advanced transmission schemes and usage of bandwidth resources. Another very promising approach is the introduction of multiple antennas at one or both ends of a link to exploit the spatial dimension of signal transmission for improved link quality and enhanced system capacity. Smart antenna concepts are extensively discussed in this context. The application of concepts with multiple antennas necessitates the introduction of more advanced and computational expensive transmitter and receiver structures, where space-time (ST) processing techniques are required to carry out spatial and temporal information processing jointly. This article introduces a new ST processing concept to enable reduced dimension ST receiver signal processing. The signal dimension can be considerably reduced compared to the number of antennas by exploiting spatial correlation properties of the received antenna signals. The associated signal transformation applies the concept of the Karhunen-Loève transformation (KLT). A great advantage of the proposed ST processing concept over traditional multiple antenna approaches is the insensitivity of the algorithms to the antenna characteristics and antenna spacing, which allows the use of low-cost antennas. Another significant advantage of the proposed concept is more robust channel estimation due to spatial dimension reduction and the resulting limitation of estimation parameters.

INTRODUCTION

One of the ultimate goals in the design of new wireless communication systems is increasing system capacity. This includes support of a growing number of users and at the same time provision of high and variable data rates for every single link. The aim to enable high data rates for a single user could be achieved by allocating more bandwidth. However, this approach will reduce the number of possible users who can be active within a system at a given available bandwidth. Another possibility is the application of spectrally more efficient modulation schemes, which enable the transmission of a greater overall amount of user data within a given time for a given bandwidth. This approach provides the means to support high-data-rate links without decreasing the number of active users. The drawback of such advanced modulation schemes is their lack of robustness against channel distortion and noise, which will result in an increased amount of transmission errors.

Therefore, the question arises of how those channel imperfections can be compensated to reduce the number of transmission errors. In this context the introduction of multiple antennas for data transmission and/or reception can be of great advantage. If an antenna array is introduced at one end of a wireless link, the given data stream will be transmitted over multiple, statistically more or less independent wireless channels. The application of multiple antennas at the transmitter and receiver results in the possibility to transmit multiple independent data streams simultaneously, where the number of independent streams is limited by the minimum of the number of antennas at the transmitter and receiver. Space-time (ST) coding concepts are frequently proposed for systems with multiple antennas at both ends of the communication link [1].

In this article we limit our attention to the case where multiple antennas are applied at the receiver only. There are several reasons why this case is interesting to analyze. First, there is no need to feed information back to the transmitter as in many multiple transmit antenna systems that apply transmit beamforming, advanced transmit diversity schemes, or other channel predistortion techniques to improve the signal reception at the receiver. Second, in wireless LANs one could imagine mobile terminals such as laptop computers that could be equipped with multiple antennas and some advanced signal processing functionality, which would then allow

1536-1284/02/$17.00 © 2002 IEEE

higher transmission rates toward the mobile terminal or significant range extension, even if only a simple access point with a single antenna is available. Finally, since the proposed concept is a pure preprocessing stage in the receiver chain, it does not prevent the application of multiple transmit antennas or multiple data streams as applied in ST coding schemes. In the latter case, the data selection stage discussed later in this article would require additional knowledge such as dedicated training symbols for an appropriate signal component selection.

A basic problem that arises with the introduction of antenna arrays is the increasing complexity of the radio front-end and the associated digital signal processing unit. Spatial and temporal processing of the received antenna signals are now required instead of a pure temporal equalization of the received signals from only one antenna. The spatial and temporal processing stages can be combined into a joint ST processing unit or treated separately. In the latter approach the antenna signals are commonly combined within a spatial processing stage, which is often implemented as a beamformer, smart antenna processor [2, 3], or spatial diversity combiner [4, 5]. This can often be viewed as a spatial preprocessing stage in front of a classical receiver. The first approach of joint ST processing is the more general approach, which can be adapted to varying scenarios to yield the expected performance improvements. These improvements heavily depend on the wireless channel characteristics, especially on the degree of spatial correlation, which determines the ability of the receiver to compensate for channel distortions such as fast fading effects. The most important improvements that can be achieved by applying antenna array concepts are:

- The signal-to-noise ratio (SNR) of the desired signal can be improved proportionally to the number of antenna elements through coherent superposition of the antenna signals.
- Since the weighted combination of the antenna signals can be viewed as a spatial filtering operation, undesired signals can be suppressed spatially. This yields an improved signal-to-interference-and-noise ratio (SINR).
- Several users can be active simultaneously in the same frequency band if they are separable otherwise. When multiple antennas are applied, different users can be separated spatially. This concept of space-division multiple access (SDMA) enables the improvement of user and system capacity.

Besides these potential benefits multi-antenna systems can efficiently be used to minimize fading effects by exploiting spatial diversity. The basic principle of this concept can be described as follows: if the channel distortions associated with the different antennas are uncorrelated, there is a high probability that if the signal strength heavily drops at one antenna element, there is a larger signal strength at other antenna elements. This effect will in turn lead to a smaller error rate with a suitable antenna signal combination. The exploitation of the spatial correlation properties, which heavily influence the possible spatial diversity gain, leads us to a new approach for the integration of spatial and temporal processing units into a reduced dimension ST processor as described in [6]. This approach includes two stages, which are discussed from a conceptual point of view in this article. In the first stage the received signal components are decorrelated using a linear transformation. Thereafter, a second stage removes transformed signal components that are not relevant for subsequent signal processing units. Additionally, we address the implications of the proposed spatial dimension reduction approach for the succeeding reduced dimension ST equalizer stage. The performance of the reduced dimension ST receiver concept, which uses the Karhunen-Loève transformation (KLT) [7] as a key component, is also analyzed. The KLT can be described as a signal adapted coordinate transformation, which in our context applies the eigenvector matrix of the spatial correlation matrix of the received antenna signals as an orthonormal projection matrix to generate output signals with maximum decorrelation. The transformed signal vector contains uncorrelated components with signal energies corresponding to the eigenvalues of the spatial correlation matrix. This implies that the eigenvalue strength can be used as a measure of the significance of the transformed signal components for succeeding signal processing stages, as is shown in this article. Removing components corresponding to the smallest eigenvalues results in the desired dimension reduction with the lowest possible signal energy loss. The KLT will be described in more detail later. For a detailed analysis of the properties of the spatial dimension reduction stage for flat fading channels see [6].

Receiver Architectures

Typical indoor propagation scenarios are commonly characterized by multipath propagation of the transmitted signal on its way to the receiver antennas. These multipaths are caused by reflections and scattering effects at objects in the propagation environment, which is exemplified in Fig. 1a for a typical indoor scenario. Those propagation effects will result in a temporal spread of the received signal due to the different propagation delays and in angular signal spread depending on the location of the reflectors and scatterers. The most general approach for a receiver equipped with multiple antennas to compensate for those propagation effects is the application of a fully armed joint space-time equalizer [8] as depicted in Fig. 2a, where M is the number of antennas and N is the number of temporal equalizer taps. The main advantage of such ST processors is joint spatial and temporal parameter (channel) estimation and equalization in contrast to independent treatment of spatial and temporal equalizer stages. At the same time, joint ST approaches face the problem of a high number of free parameters to be estimated. The full dimension ST processor with M antenna elements and N temporal equalizer taps requires the estimation of $M \times N$ parameters. Besides the computational load of this estimation process, numerical problems may arise for certain channel conditions. Assuming the case where only one wave-

The main advantage of such ST processors is joint spatial and temporal parameter (channel) estimation and equalization in contrast to independent treatment of spatial and temporal equalizer stages.

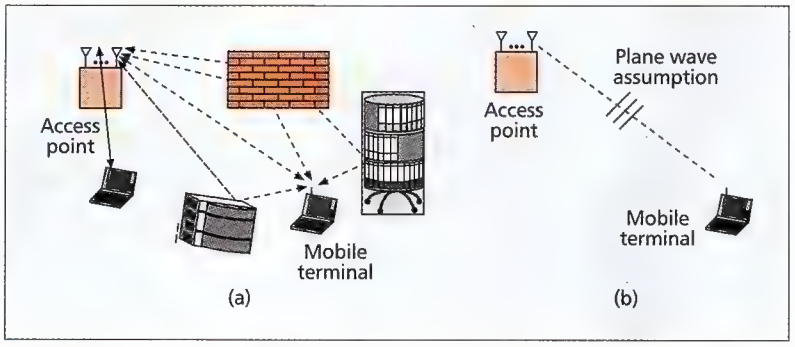


Figure 1. *a) A typical indoor communication scenario; b) an adaptive beamforming (line of sight) scenario.*

front propagates across the M receive antennas as depicted in Fig. 1b, the resulting received signals and their respective channels are almost identical up to a phase shift, which corresponds to the incidence angle of the wavefront and the array geometry, and some decorrelation caused by additive noise. The associated channel impulse responses for all receiver chains are in this case linearly dependent. If we introduce a channel matrix $\mathbf{H}(t)$ containing the M channel impulse responses of length N as rows, this channel matrix will be close to singular [6, 9].

Therefore, in environments with only one strong propagation path or small angular spread, the application of a separate spatial processor (beamformer, spatial diversity combiner) and a temporal equalizer as indicated in Fig. 2c might be more feasible. For this beamforming approach the optimization of M beamforming weights and estimation of N temporal equalizer coefficients are required. However, this approach will limit the applicability of the receiver, since it is adapted to the above mentioned propagation conditions.

Looking at the previously discussed equalizer structures it becomes obvious that there are several degrees of freedom for the implementation of a structure that can be placed (in terms of complexity and flexibility) between the spatial combining approach and the fully armed ST processor, which can be considered limiting cases. An important reason for this is that even in multipath scenarios with considerable angular and temporal spreads there exists partial correlation between the received signals at the M antenna elements. This correlation can be removed in a preprocessing or transformation stage. Additionally, this effect enables the reduction of the signal space and a reduced dimension ST processor, as shown in Fig. 2b with $D_s \leq M$ as reduced spatial dimension. This in turn considerably enhances the channel estimation when the transformation results in improved SNRs for the D_s transformed signal components.

Reduced Dimension Space-Time Receiver Concept

Considering the basic principles and the structure of the proposed ST receiver, the main challenges in the receiver design are twofold: first, we need to find an appropriate algorithm to decorrelate the antenna signals; second, we have to determine a selection criterion to remove components that do not significantly contribute to receiver performance. The first step will be referred to as the *transformation stage* and the second as the *selection stage*. The selection stage is the means to reduce the signal dimension and is therefore the key to enabling reduced dimension ST equalization. The principle is depicted in Fig. 3. Clearly, the design of the transformation and selection stages should be done jointly. The transformation stage should move the desired information into as few transformed signal components as possible to allow the selection stage to efficiently reduce the dimensionality.

Spatial Correlation Properties

The spatial correlation properties of the received antenna signals determine the possible degree of the dimension reduction. Therefore, we introduce the spatial correlation matrix $\mathbf{R}_{\mathbf{x}}^{S}$ of the received signal vector $\mathbf{x}(t)$. To simplify the interpretation, we assume zero mean and independent identically distributed data symbols as well as spatially and temporally white noise, which is uncorrelated with the data symbols. Then we can write the spatial correlation matrix as

$$\mathbf{R}_{\mathbf{x}}^{S} = \mathcal{E}\{\mathbf{x}(t)\mathbf{x}(t)^{H}\} = \mathcal{E}\{\mathbf{H}\mathbf{H}^{H}\} + \mathbf{R}_{\mathbf{n}}^{S} \qquad (1)$$

with $\mathbf{R}_{\mathbf{n}}^{S} = \sigma_n^2\mathbf{I}$ as diagonal matrix containing identical noise variance terms σ_n^2 for all antenna elements. The matrix $\mathbf{H}$ of dimension $M \times N$ describes the transmission channel for M receive antennas and a temporal channel response with

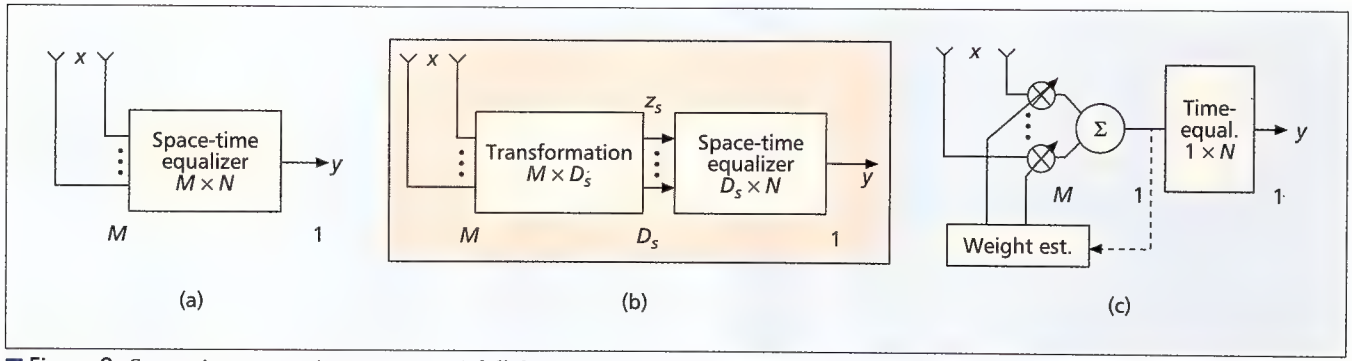


Figure 2. *Space-time processing concepts: a) full dimension ST processing; b) reduced dimension ST processing; c) spatial combining followed by temporal equalization, with* N = *number of antennas,* N = *number of temporal equalizer taps,* D_s = *number of spatial dimensions after transformation,* **x** *as received signal vector,* y *as equalized signal.*

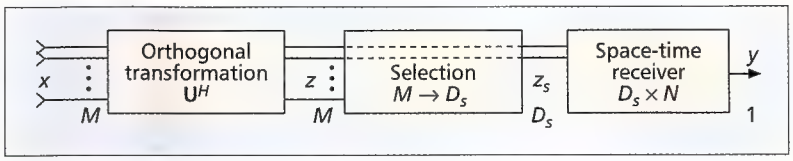
a duration limited to N symbol periods. The properties of the transmission channel are influenced by the propagation scenario, namely the temporal and angular signal spread introduced by obstacles in the environment, the number of significant propagation paths, and their respective propagation loss, but also by filters in the transmission chain and antenna characteristics such as the geometry of the antenna array. These parameters determine the spatial correlation properties. Recalling the beamforming scenario depicted in Fig. 1b and assuming, that the antenna elements are spaced sufficiently close and that no noise is present, the signals at the receive antennas are basically phase shifted copies of each other due to the signal propagation across the antenna array. This results in M almost perfectly correlated channels and hence in a channel matrix $\mathbf{H}$ with linearly dependent rows. Therefore, the $\mathbf{H}$ and $\mathbf{R}_{\mathbf{x}}^{S}$ matrices will have a rank of 1 in this special case. More generally, scenarios with fewer multipath components than antennas M will result in spatial correlation matrices $\mathbf{R}_{\mathbf{x}}^{S}$ that are rank deficient or close to rank deficient in noisy environments. Although this is not true for rich scattering scenarios, which typically occur in wireless indoor communication scenarios as depicted in Fig. 1a, even in these cases the received signals and the associated channels are partially correlated [10].

To further investigate these spatial correlation properties in conjunction with the dimension reduction approach, we will introduce the singular value decomposition (SVD) [7]. This decomposition is a powerful tool for a numerically stable estimation of the eigenvalues and eigenvectors of hermitian matrices, especially if they are ill conditioned or close to singular. The hermitian correlation matrix $\mathbf{R}_{\mathbf{x}}^{S}$ can be decomposed into the form

$$\mathbf{R}_{\mathbf{x}}^{S} = \mathbf{U}\Lambda\mathbf{U}^{H} \tag{2}$$

with $\mathbf{U} = [\mathbf{u}_0\ \mathbf{u}_1\ \ldots\ \mathbf{u}_{M-1}]$ as the lefthand side singular vector matrix, or equivalently eigenvector matrix of $\mathbf{R}_{\mathbf{x}}^{S}$ containing the M eigenvectors $\mathbf{u}_i$ and Λ as a diagonal matrix containing the singular values or eigenvalues sorted in descending order, $\Lambda = \text{diag}\,[\lambda_0^2\ \lambda_1^2\ \ldots\ \lambda_{M-1}^2]$. Note that the eigenvalues are determined by the received signal properties and the noise variance, $\lambda_i^2 = \sigma_i^2 + \sigma_n^2$, as can be seen from Eq. 1.

In order to clarify the influence of different propagation scenarios on the eigenvalue distribution, the cumulative distribution functions of the normalized eigenvalues $\tilde{\lambda}_i^2 = \lambda_i^2 / \Sigma_{m=0}^{M-1} \lambda_m^2$ are shown in Fig. 4 for two different propagation scenarios for a noise-free case. We define a uniform path delay distribution, which is limited to a fixed maximum path delay of four times the symbol duration to model the path delays from scattered and reflected paths relative to the first arriving propagation path. Furthermore, a uniform angular distribution of the impinging multipath components is assumed. Two different beamwidth values are selected for the angular distribution visible at the receive antennas relative to a main direction of 0°. In the first scenario this beamwidth is set to $\Delta_\theta = \pm 4°$, which implies strong spatial signal correlation due to almost identical incidence angles, and to $\Delta_\theta = \pm 70°$ for low antenna signal correlation. A large number of channel realizations is simulated by repeatedly drawing path delays and angles from random processes applying the above mentioned distributions. The small angular beamwidth in the first scenario yields an eigenvalue distribution as depicted in Fig. 4a. There is with very high probability only one strong eigenvalue $\tilde{\lambda}_0^2$ corresponding to the main direction, the second eigenvalue is already much smaller. For large angular spread values the situation turns out to be quite different as indicated by Fig. 4b. There exist several relatively strong eigenvalues. This indicates that the different antenna signals are rather uncorrelated due to different superposition of the multipath signal components at the single antenna elements. The next paragraphs will show how these eigenvalue properties can be used for spatial dimension reduction.



■ Figure 3. *Reduced dimension ST receiver concept.*

The Transformation Stage

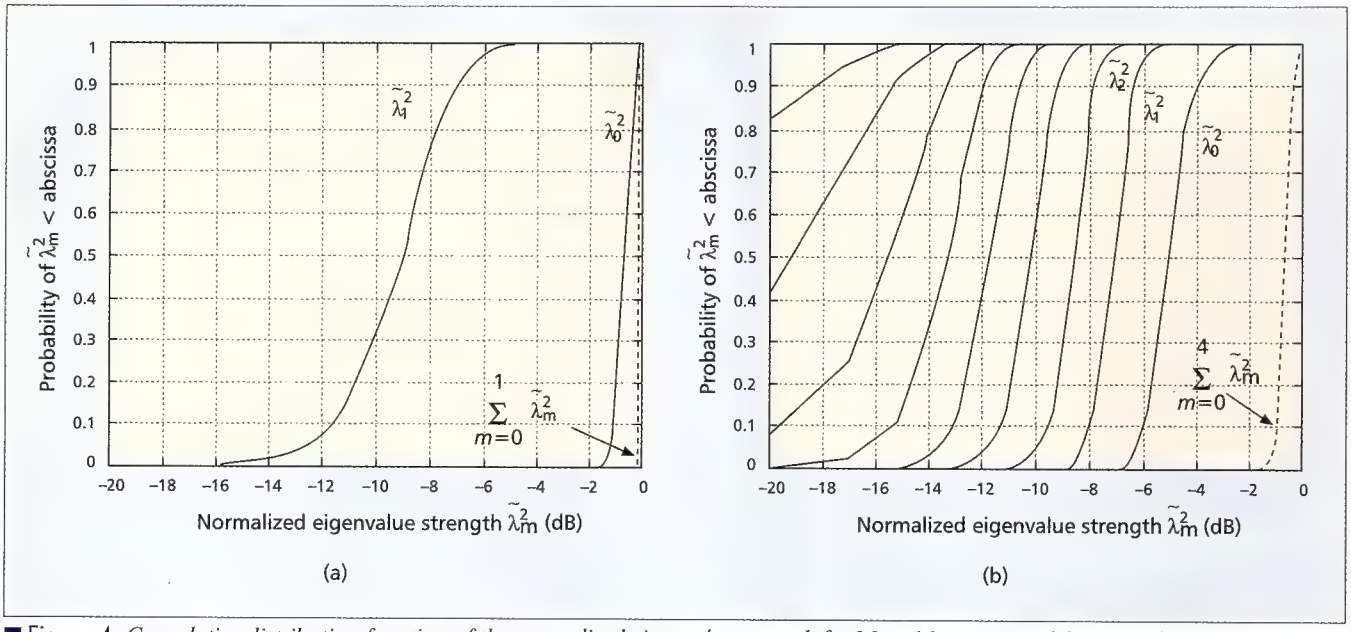
In order to be able to efficiently reduce the spatial signal dimension, we need to find an orthogonal transformation, which decorrelates the received antenna signals and, at the same time, concentrates the signal energy within as few transformed signal components as possible. The KLT [7] in this context plays a key role among the orthogonal transformations, since it is signal adapted by directly using the signal correlation properties as described previously. If the M-dimensional signal vector $\mathbf{x}(t)$ results from a wide sense stationary vector process with zero mean and correlation matrix $\mathbf{R}_{\mathbf{x}}^{S}$, this vector can be expanded as a linear combination of the eigenvectors $\mathbf{u}_i$ of $\mathbf{R}_{\mathbf{x}}^{S}$,

$$\mathbf{x}(t) = \sum_{i=0}^{M-1} z_i(t)\mathbf{u}_i = \mathbf{U}\mathbf{z}(t). \tag{3}$$

The associated coefficients $z_i(t)$ are zero mean and uncorrelated random variables, which can be represented in vector notation as

$$\mathbf{z}(t) = \mathbf{U}^H\mathbf{x}(t). \tag{4}$$

This equation exactly describes the desired transformation stage as was shown in Fig. 3. The transformation matrix $\mathbf{U}$ contains the eigenvectors of the spatial correlation matrix $\mathbf{R}_{\mathbf{x}}^{S}$ which makes the KLT a signal-adapted transformation. The resulting components in the transformed signal vector $\mathbf{z}(t)$ are uncorrelated. Furthermore, the signal energy of the transformed vector components $z_i(t)$ is equal to the respective eigenvalues $\lambda_i^2 = \mathcal{E}\{|z_i(t)|^2\}$ of the spatial correlation matrix [7]. This property results in transformed components $z_i(t)$ with unequal mean signal energy determined by the eigenvalues λ_i^2 with maximum energy in the first component, since the eigenvalues are assumed to be sorted in descending order. This fact, that the signal energy is con-



■ Figure 4. *Cumulative distribution function of the normalized eigenvalue strength for* M = *16 antennas, delay spread* $\sigma_t = 4T_{sym}$: *a) small angular beamwidth* $\Delta_\theta = \pm 4°$; *b) large angular beamwidth* $\Delta_\theta = \pm 70°$.

centrated in the first transformed components $z_i(t)$, is the key to the dimension reduction with minimum signal energy loss.

Using this property, it turns out that there exists a great potential to reduce the spatial signal dimension for scenarios with eigenvalue distributions as depicted in Fig. 4a. On the other hand, the dimension reduction potential is limited for eigenvalue distributions as shown in Fig. 4b corresponding to scenarios with large temporal and angular spread. However, the dashed line shows the sum of the signal energy or eigenvalues of the five strongest transformed signal components, which clearly indicates, that in this example the selection of the five strongest of 16 components will result in a loss of signal energy less than or about 1 dB. Consequently, also this scenario provides a considerable dimension reduction potential.

The Dimension Selection Stage

Having transformed the received signals we need to smartly select only those transformed signal components $z_i(t)$, which are essential for the subsequent ST processing stage. As discussed earlier in this section, in some cases the spatial correlation matrix will be rank deficient or close to rank deficient. This is especially true if there exist less multipath components than antenna elements. In such situations the space spanned by the eigenvectors $\mathbf{u}_i$ is larger than the actual signal space. Stated differently, there exist eigenvectors that exclusively correspond to noise components, with the respective eigenvalues being equal to the noise variance σ_n^2. Removing those components that belong to the so-called noise subspace will reduce the overall noise variance of the received signal and therefore enable improved receiver performance. There exist several powerful algorithms to estimate the dimension of the signal and noise subspaces, which are known as information theoretic criteria [11].

However, in many wireless communications scenarios with rich multipath scattering and a rather limited number of receive antennas, the method described above will not be appropriate, when the number of signal components exceeds the number of antennas. Therefore, a different approach is suggested here, which is known as low-rank modeling [12]. The basic idea is to trade the modeling error introduced by removing signal components with weak eigenvalues λ_i^2 for the noise variance saved by removing noisy components.

From Eq. 3 it can be seen that the received signal vector can be expanded as a linear combination of the eigenvectors $\mathbf{u}_i$. If components are removed from this sum, an approximation error $\varepsilon_{\tilde{\mathbf{r}}}$ is introduced, which is referred to as *bias*. This bias heavily depends on the eigenvalue distribution. The mean-squared error (MSE) introduced by removing a signal component $z_i(t)$ is equal to the corresponding eigenvalue λ_i^2 or signal energy [9, 12]. Consequently, the strength of the bias introduced by removing components with small eigenvalues will depend on the ratio between strong and weak eigenvalues. A large ratio between strong and weak eigenvalues will result in a small bias and vice versa.

On the other hand, with the assumption of equal noise variance for all received antenna signals this noise variance will remain unchanged after the transformation stage. Therefore, removing transformed signal components will also reduce the overall noise variance and the respective MSE $\varepsilon_{\tilde{\mathbf{n}}}$. This effect will in turn decrease the MSE of the approximated receive signal. Hence, the removal of transformed signal components can be viewed as a bias-variance trade-off.

Comparing the two effects it can be shown that the low rank model [12] as well as the transformed received signal will be improved in the MSE sense, when the neglected signal energy is less than the removed noise variance,

$$\sum_{i=D_s}^{M-1} \sigma_i^2 \le (M - D_s)\sigma_n^2, \qquad (5)$$

where D_s is the truncated spatial signal dimension and $\sigma_i^2 = \lambda_i^2 - \sigma_n^2$ is the ith small eigenvalue of the transformed component $z_i(t)$ revised by the noise variance σ_n^2. An example of the bias-variance trade-off is shown in Fig. 5 for $M = 8$ antennas, a mean input SNR per antenna of $SNR_i = 3$ dB, and assuming rapidly decaying eigenvalues as indicated by the dashed line. This curve shows the MSE $\varepsilon_{\hat{r}}$ introduced when approximating the received signal by D_s instead of M components (cf. Eq. 3). The dashed-dotted line visualizes the error $\varepsilon_{\hat{n}}$ caused by the overall noise variance associated with the number of components D_s. Finally, the solid curve shows the resulting MSE $\varepsilon_{\hat{x}}$ introduced by the dimension reduction, which in this example reaches its minimum value for $D_s = 2$ transformed components. It can be observed from Fig. 5 that the ratios between the eigenvalues σ_i^2 strongly influence the optimum dimension D_s. However, it can be shown that for most scenarios the number of strong eigenvalues is very limited. For antenna configurations of 8 and 16 antennas the spatial dimension can frequently be reduced to $D_s = 1 \ldots 3$ [9, 10].

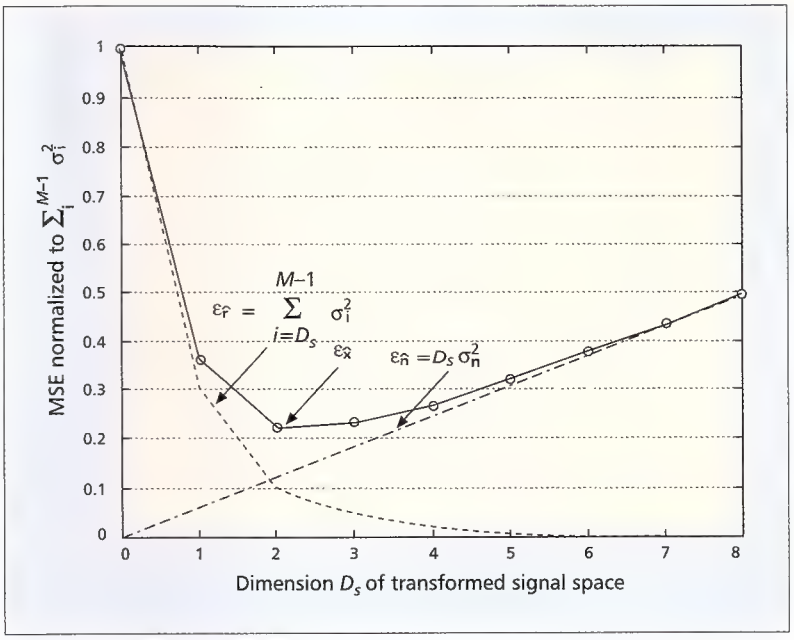
If we compare these results with the rank deficiency approach discussed earlier, the advantages become clear. In the example discussed in Fig. 5 no rank deficiency exists. This means that with traditional methods to determine the signal subspace dimension, no reduction of the spatial dimension would result. In contrast, the application of the bias-variance trade-off provides considerable dimension reduction potential. Moreover, at the same time the MSE of the transformed vector signal is minimized. This effect will be further discussed in the following subsection.

The selection process discussed here does not depend on additional a priori information. However, depending on the actual system it might be essential to exploit additional user-specific information, either in front of the transformation stage or for the selection stage.

The Reduced Dimension Space-Time Equalizer Stage

Finally, we need to investigate the implications of the spatial dimension reduction for the succeeding ST equalizer. There exist various ST extensions for conventional equalizer structures. These include ST versions of the minimum mean squared error (MMSE) equalizer, the decision feedback equalizer (DFE), and the maximum-likelihood sequence estimation (MLSE) equalizer [13–15].

We will here focus our attention on the ST-MLSE equalizer, which can efficiently be implemented as an extension of the well-known Viterbi equalizer [15]. One key element of the Viterbi equalizer is the branch metric (BM) computation, which is performed for every symbol. We will show here how the reduced dimension approach influences this BM computation process and the receiver performance. The ST extension of the Viterbi equalizer is as follows: the BMs are independently computed for all M received signal branches; the results are then summed to the overall BM accordingly. Reducing the dimension of the received signals to D_s results in a linear decrease of the computational load.

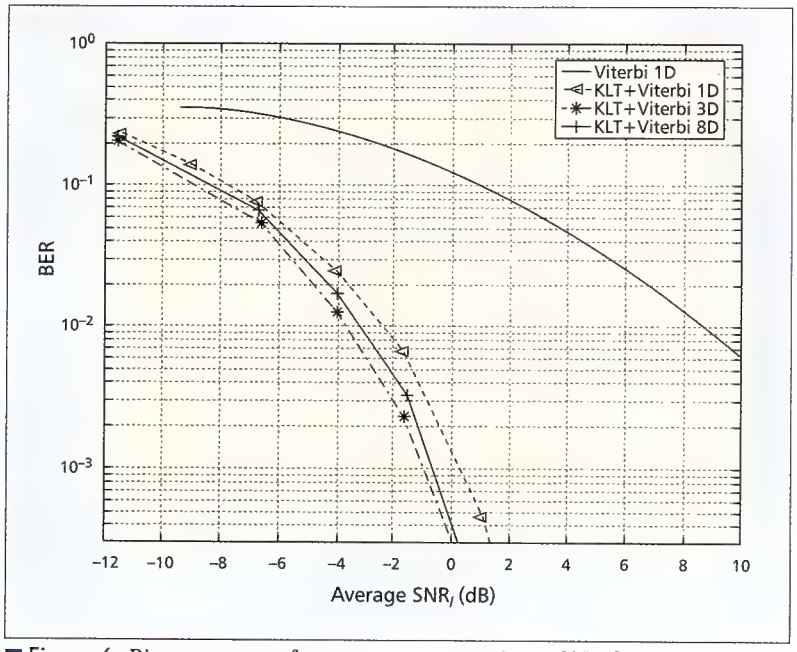


■ Figure 5. *Bias-variance trade-off for* M = *8 antennas, mean input SNR of 3 dB, and normalized eigenvalue distribution diag(Σ) =[0.7, 0.2, 0.05, 0.03, 0.01, 0.005, 0.003, 0.002]*^{T}.

Moreover, the BM computation requires estimates of the channel parameters as input. Improved channel estimation therefore directly influences equalizer performance. Using the reduced dimension signal stream as input, the channel estimation problem can be considerably simplified. Instead of $M \times N$ parameters, now only $D_s \times N$ channel coefficients need to be estimated. Furthermore, as described above, the dimension reduction to the optimum dimension D_s also reduces the MSE and the overall noise variance of the transformed signal. Hence, the channel estimation will become more reliable. The underlying reason for this effect is that besides a short training sequence, additional knowledge about the received signal contained in the spatial correlation matrix is exploited for the channel estimation.

Results

Following the performance of the proposed reduced dimension ST receiver will be analyzed. In the discussed example BPSK modulated data were transmitted from a transmitter with one antenna to a receiver with $M = 8$ antenna elements. The channel characteristics were modeled using a tapped delay line model with 12 taps, which is characterized by Rayleigh fading components and rather high temporal spread. Additionally, an angular value was assigned as spatial information to every multipath tap. The angles were drawn from a uniform distribution with an angular beamwidth of ±70°. The spatial correlation matrix was estimated for every data block of length 2093 including a training sequence of length 93. The SVD was performed



■ Figure 6. *Bit error rate performance vs. average input SNR for a reduced dimension ST-Viterbi equalizer.*

for the spatial correlation matrix estimate. The receiver performance was investigated for various ST receiver dimensions D_s. To limit the complexity of the Viterbi equalizer, the number of temporal channel taps was set to $N = 6$. The achievable BERs are shown in Fig. 6 and plotted over the average input SNR per antenna in dB.

The solid rightmost curve shows a reference simulation for a single antenna element. When this curve is compared with the other curves, which present simulation results for eight antenna elements, considerable gains of ≥ 9 dB can be identified for the multi-antenna approach. This gain is determined by two effects. First, the usage of M antenna elements results in a SNR gain of $10\log(M) = 9$ dB. This is a well-known result from beamforming approaches [2], which also holds for the proposed dimension reduction technique. Second, a spatial diversity gain is achieved by combining multiple antenna signals. This means that the probability of deep fades in the transformed signal components is considerably reduced, resulting in a reduced variance of the SNR at the output of the dimension reduction stage. The achievable diversity gain strongly depends on the degree of correlation between the different received signals. If we recall the scenario shown in Fig. 1b, we can draw an important conclusion. For cases where only one strong path and no scattering is present, the proposed procedure yields the same results as a classical beamformer. The dimension can be reduced to $D_s = 1$, since a single strong path corresponds to a single strong signal eigenvalue. Due to the total signal correlation no spatial diversity can be achieved. If, on the other hand, we assume scattering around the receiver, but with very limited temporal spread, which results in a frequency flat channel, we also end up with only one strong eigenvalue. However, in this case the antenna signals are rather uncorrelated, resulting in an additional spatial diversity gain. For statistically independent spatial channels the proposed dimension reduction approach therefore also includes the maximal ratio (diversity) combining (MRC) scheme. Hence, we can conclude that the discussed approach brings together the two concepts of beamforming and MRC as special cases, which are usually considered as independent concepts. Another important statement is that, using the reduced dimension ST receiver concept, the antenna geometry does not play an important role as it does for beamforming concepts, since the antennas are simply considered as "independent" information sources.

Comparing the three left curves in Fig. 6 in more detail, it turns out that the influence of the number of selected components on the achievable gain for the given channel conditions is limited. The gain difference between the dimension reduction to one (– – ◁ – –) and the optimum dimension is less than 2 dB for the considered SNR range. For very low SNR values the channel estimation is very noisy. In this case the dimension reduction to $D_s = 1$ yields the best results. For increasing SNR values the loss caused by neglecting information bearing components becomes dominant compared to the noise variance reduction. When comparing the curves for dimension reduction to $D_s = 3$ (– · ∗ · –) and to 8 (—+—) components, the full dimension case performs worse in the entire SNR range. This indicates that for the given channel model the signal energy will be concentrated in the first transformed components, the components corresponding to small eigenvalues basically contain only noise. For the optimum dimension selection case, applying the bias-variance trade-off the respective BER curve is determined by the minimum of the array of curves for all possible dimensions D_s. However, as indicated before for 8 and 16 receive antennas, the number of strong eigenvalues was limited to $D_s = 1 \,..\, 3$ for all considered channel conditions. This suggests that the space-time equalizer would require not more than three spatial dimensions to perform well for a great variety of investigated channels, which considerably limits the equalizer complexity. The application of the bias-variance trade-off could then indicate additional spatial dimension reduction potential to save processing power by deactivating unnecessary spatial equalizer processing stages.

The results discussed in this section for BPSK will also hold for higher order modulation schemes (M-PSK, M-QAM) commonly applied in wireless LANs, since for independent identically distributed data with zero mean the dimension reduction stage does not change the properties of the transmitted signal waveforms. Moreover, those higher order modulation schemes will profit even more from the improved channel estimation, especially in the low SNR region, since they are much more sensitive to channel estimation errors.

Conclusion

Emerging wireless services with high and variable data rate demands and the increasing number of mobile users require efficient usage of available resources. The exploitation of the spa-

tial dimension by applying multiple transmit/receive antennas is a promising approach to increasing the required system performance. In this article a new reduced dimension space-time processor concept has been introduced. Using the eigenvalue properties of the spatial correlation matrix of the received signals, the optimum reduced receiver dimension can be determined by applying a bias-variance trade-off. In contrast to many subspace approaches, this concept also works in rich multipath scenarios. It is important to note that the antenna configuration is not of major concern for the approach discussed here as compared to beamforming algorithms, whose performance is highly dependent on antenna characteristics, especially when direction-of-arrival estimation techniques are involved. The reason for this property is the treatment of the antenna signals as independent signal components contributing to the overall information about the received signal available at the receiver, which involves no special assumptions concerning the antenna geometry. Hence, this approach enables the usage of low cost antennas. Furthermore, antenna configurations enhancing the spatial decorrelation can be applied to improve the achievable diversity gain. Besides the dimension reduction of the ST equalizer, a significant advantage of the proposed approach is enhanced channel estimation. This issue will become even more relevant when multiple transmit antennas are involved, since in such systems the transmit power per antenna element will be decreased, resulting in a more challenging channel estimation task. When multiple transmit antennas are introduced, the proposed concept can be applied without modification if transmit beamforming is performed. For systems with multiple transmitted data streams the concept requires a modification in the selection stage. The usage of dedicated training symbols would then be required to select the relevant signal components for all independent data streams, which implies that the minimum dimension would in this case be equal to the number of independent data streams.

References

[1] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Info. Theory*, vol. 44, no. 2, Mar 1998, pp. 744–64.

[2] J. Litva and T. Kwok-Yeung Lo, *Digital Beam-forming in Wireless Communications*, Artech House, 1996.

[3] H. Krim and M. Viberg, "Two Decades of Array Signal Processing Research," *IEEE Sig. Proc.*, July 1996, pp. 67–94.

[4] W. C. Jakes, Ed., *Microwave Mobile Comun.*, IEEE Press, 1984.

[5] J. H. Winters, "Optimum combining in digital mobile radio with cochannel interference," *IEEE Trans. Vehic. Techn.*, vol. 33, no. 3, Aug. 1984, pp. 144–55.

[6] J. Jelitto, M. Bronzel, and G. P, Fettweis, "From Smart Antennas to Reduced Dimension Space-Time Processing," *FREQUENZ, J. of Telecommun.*, vol. 55, nos. 5–6, May/June 2001, pp. 165–170.

[7] S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice Hall, 1996.

[8] A. Paulraj and C. B. Papadias, "Space-Time Processing for Wireless Communications," *IEEE Sig. Proc.*, vol. 14, no. 6, Nov. 1997, pp. 49–83.

[9] J. Jelitto, M. Bronzel, and G. P. Fettweis, "Reduced complexity space-time optimum processing," *Proc. 10th Virginia Tech/MPRG Symp. Wireless Pers. Commun.*, Blacksburg, VA, USA, 14–16, June 2000, pp. 167–78.

[10] J. Jelitto, "Dimensionsreduktion des Empfangsraumes von Systemen mit mehreren Empfangsantennen und deren Anwendung in der Raum-Zeit-Verarbeitung," *Fortschrittberichte VDI*, vol. 10, no. 693, VDI Verlag, 2002.

[11] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech and Sig. Proc.*, vol. 33, no. 2, April 1985, pp. 387–92.

[12] L. L, *Scharf, Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley, 1990.

[13] J.-W. Liang, "Interference Reduction and Equalization with Space-Time Processing in TDMA Cellular Networks," Ph.D. thesis, Stanford Univ., June 1998.

[14] E. Lindskog, "Space-Time Processing and Equalization for Wireless Communications," Ph.D. thesis, Uppsala Univ., 1999.

[15] P. Jung, M. Naßhan, and Y. Ma, "Comparison of optimum detectors for coherent receiver antenna diversity in GSM type mobile radio systems," *Proc. PIMRC '93*, Yokohama, Japan, 1993, pp. 54–58.

Biographies

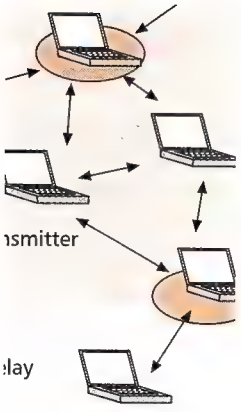
GERHARD FETTWEIS [S'82-M'90-SM'98] (fettweis@ifn.et.tu-dresden.de) received his M.Sc./Dipl.-Ing, and Ph.D. degrees in electrical engineering from Aachen University of Technology (RWTH), Germany, in 1986 and 1990, respectively, From 1990 to 1991 he was a visiting scientist at the IBM Almaden Research Center, San Jose, California, working on signal processing for disk drives, From 1991 to 1994 he was a scientist with TCSI, Berkeley, California, responsible for signal processor developments for mobile phones, Since September 1994 he holds the Mannesmann Mobilfunk Chair for Mobile Communications Systems at Dresden University of Technology, Germany, He is an elected member of the SSC Society's Administrative Committee and IEEE ComSoc Board of Governors, since 1999 and 1998, respectively, He has been associate editor for *IEEE Transactions on CAS II*, and now is associate editor for the *IEEE JSAC* Wireless series.

JENS JELITTO (jje@zurich.ibm.com) received his M.Sc./Dipl.-Ing and Ph.D. degree from Dresden University of Technology, Germany, in 1995 and 2001, respectively. From 1995 to 1996 he worked in the field of speech recognition at the Institute for Acoustics and Speech Communication in Dresden. In July 1996 he joined the Mannesmann Mobilfunk Chair for Mobile Communications Systems at the Dresden University of Technology, Germany, to work toward his Ph.D. degree, where his main research interests included digital signal processing, smart antennas, and spatial dimension reduction problems. In March 2001 he joined the IBM Zurich Research Laboratory, Rüschlikon, Switzerland, as a research staff member, working on digital signal processing for wireless LANs.

Emerging wireless services with high and variable data rate demands and the increasing number of mobile users require the efficient usage of available resources. The exploitation of the spatial dimension by applying multiple transmit/receive antennas is a promising approach to increase the required system performance.

ADEQUACY BETWEEN MULTIMEDIA APPLICATION REQUIREMENTS AND WIRELESS PROTOCOLS FEATURES

ANTOINE MERCIER, ECOLE CENTRALE D'ELECTRONIQUE
PASCALE MINET, INRIA
LAURENT GEORGE AND GILLES MERCIER, LIIA



Recent developments in wireless communications have made it possible to provide multimedia services for mobile users through wireless devices. Multimedia applications over wireless links require that the underlying wireless network support quality of service.

ABSTRACT

Wireless networks come in force into the market of computer communications. Unfortunately, several standards coexist (e.g., IEEE 802.11, Bluetooth, HomeRF, HIPERLAN), leading to various types of products. The purpose of this article is to present an overview of those standards and to establish a comparative evaluation. We mainly focus on the medium access control protocols and their ability to support the quality of service required by multimedia applications. Those protocols generally support two modes of operation, a random access mode for asynchronous traffic and a polling mode for real-time synchronous traffic. This article analyzes each protocol's ability to support user QoS requirements.

INTRODUCTION

Recent developments in wireless communications have made it possible to provide multimedia services for mobile users through wireless devices. Multimedia applications over wireless links require that the underlying wireless network support quality of service (QoS). Current wireless standards such as IEEE 802.11, Bluetooth, HomeRF, and HIPERLAN 1 and 2 try to support multimedia services in the most appropriate way.

This article is organized as follows. Before describing these five wireless protocols, we establish several comparative criteria in the next section. We consider the traffic submitted by an application to a wireless network and express how the application's requirements are taken into account by the wireless network. Different traffic classes can be distinguished (e.g., audio/video streaming, bulk transfer, Web traffic); each class is defined by specific features (e.g. bandwidth, transmission delay, and bounded jitter). Each class of traffic requires a certain level of QoS affecting the quality of transmission perceived by the user. The QoS can be examined from two perspectives: user and network. From the user perspective, QoS requirements are based on user perception of the level of service delivered. User QoS must be translated into network QoS according to a mapping function. The network's role is to ensure that the desired QoS is met by means of management and control functions. In the third section we describe the medium access control (MAC) protocols for the five wireless standards, focusing on mechanisms that support QoS requirements according to the user point of view. We describe how user-level QoS requirements can be mapped to the low-level network. In the fourth section, a table summarizes these protocols according to the criteria given in the second section. We then study the adequacy of these five wireless standards with respect to different application constraints. Finally, we discuss some open issues for future work.

COMPARATIVE CRITERIA

We now focus on any application running on a wireless network. This application has constraints with respect to the network supporting it. We can distinguish three types of constraints: global, MAC-related, and traffic-related.

The first constraints are global, which are based on general technical considerations related to wireless networks. They essentially concern the deployment of such a network and its ability to be interconnected. Among the global constraints are:

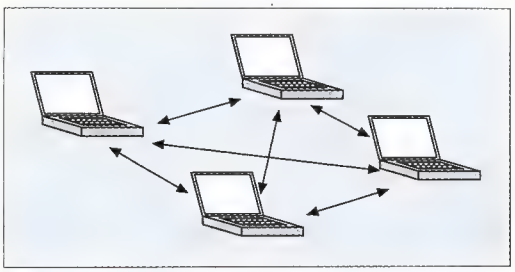
Nomadicity of devices: Some or all devices can change their location and want to keep the same virtual environment (e.g., services offered to the user); the mapping of the virtual environment to the real one depends on the physical location of the device. Such a device must be connected regardless of location.

Mobility of devices: Some or all devices can move while they are communicating. Service continuity must be provided.

Power management: Battery-operated devices must be able to save power when they are inactive. This constraint has a strong impact on protocol design.

Number of interconnected devices: This defines the maximum number of devices that

1536-1284/02/$17.00 © 2002 IEEE



■ Figure 1. *Ad hoc topology.*

can be managed in the same area under the control of a central station (if used).

Global throughput: This characterizes the maximum throughput supported by a wireless network

Frequency band: This is the range of radio frequencies used for transmission and reception, generally restricted by a country's regulations

Confidentiality: A minimum requirement is given by Wired Equivalent Privacy (WEP), where a wired LAN equivalent data confidentiality is provided to users of wireless networks.

Easy installation and maintenance

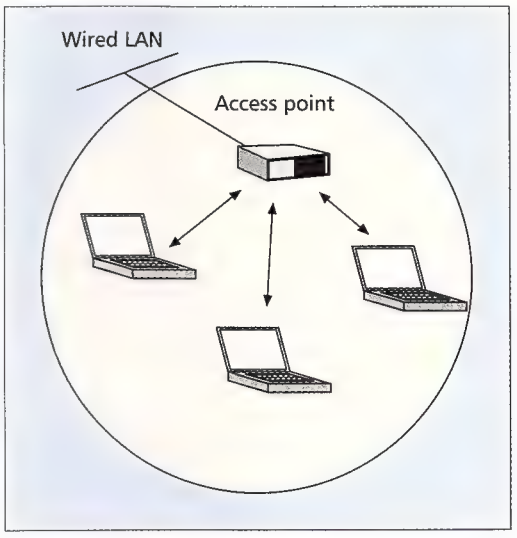
Cost effectiveness: Networking cost must be low for three reasons: first, low-cost devices; second, a high number of interconnected devices; third, the application lifetime is short.

Ability to interconnect with other networks (wired or not)

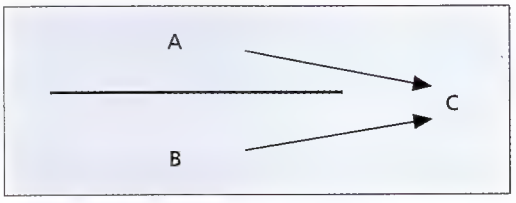
Wireless transmission provides undeniable advantages for a communication network. However, such a physical medium introduces some unique problems. A wireless MAC protocol needs to support some specific mechanisms in order to solve these problems. These include:

The need for a wired infrastructure topology: Wireless LANs (WLANs) can be built with or without an infrastructure, leading to either an infrastructure or ad hoc topology (Fig. 1). In an infrastructure topology, there is a specific device called an access point (AP). An AP not only controls the communication of devices within its radio range, but also provides access to a wired network.

Centralized/decentralized medium access: In a centralized medium access type (Fig. 2), a central entity controls the medium among the devices in its radio range (e.g., by means of a polling scheme). In this case, it ensures that all transmissions between devices under its control are contention-free. Generally, this access type is used for synchronous traffic. Moreover, with this access type, it is easier to implement the power saving mode, since the polling scheme can account for the sleeping period of battery-operated devices. In a decentralized medium access type, each device is independent and competes for medium access. Random access schemes are generally used. In a wireless environment, a device cannot detect a collision while transmitting: the transmitting power is much greater than the receiving power. This is why a collision avoidance scheme is used. Examples include carrier sense multiple access with collision avoidance (CSMA/CA) and elimination yield-non-preemptive priority multiple access (EY-NPMA). An infrastructure topology may use centralized medium access, where the control entity is the AP. An ad hoc infrastructure generally uses a decentralized medium access; however, it can also use a centralized one, where a central entity is elected.



■ Figure 2. *Centralized topology.*



■ Figure 3. *The hidden node problem.*

Centralized communication: If the medium access scheme needs a centralized communication, all traffic must be transmitted through a central entity. Each transmission is performed twice: from (or toward) the central entity in downlink direction (or uplink direction).

Communication with out-of-range devices: Since radio coverage is limited, routing solutions may need to be implemented in a WLAN. In an infrastructure topology, the wired network performs routing between devices using different APs. In an ad hoc topology, routing can be done in the WLAN itself without resorting to a wired network.

The hidden node problem: This problem (Fig. 3) occurs when two devices, A and B, that cannot hear each other transmit simultaneously to a third device, C, hearing both A and B. This generic problem in radio transmission induces loss of efficiency in terms of delay and bandwidth.

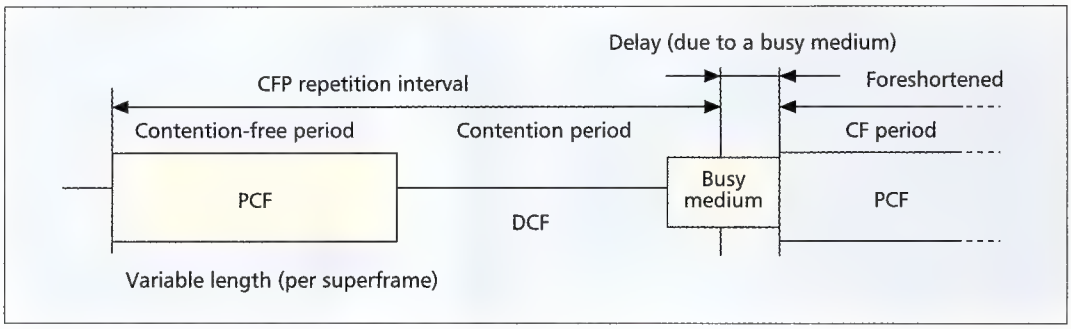
All these constraints have a major impact on MAC protocol design.

An application generates different types of traffic on the network and expresses constraints with regard to this traffic. Among these constraints, we can cite:

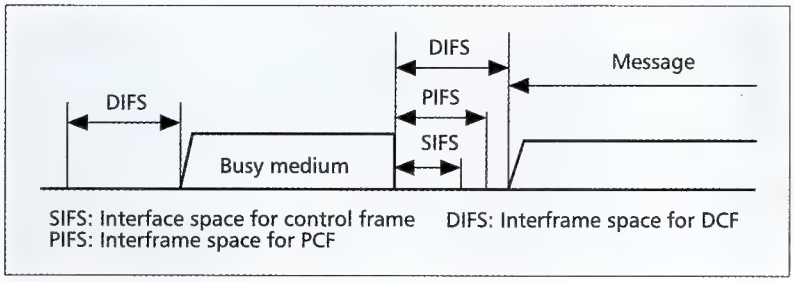
Coexistence of different types of traffic: An application can be characterized from a network point of view as a set of weighted elementary traffic. The weight and services offered to each traffic type depends on the application.

Dynamics of the traffic distribution: A wireless environment is subject to frequent changes. Consequently, the way the traffic is distributed over the devices changes in space and time: for

An infrastructure topology may use a centralized medium access, where the control entity is the AP. An ad-hoc infrastructure generally uses a decentralized medium access; however, it can also use a centralized one, where a central entity is elected.



■ Figure 4. *The basic MAC frame structure in IEEE 802.11.*



■ Figure 5. *Interframe spaces in IEEE 802.11.*

instance, devices previously in communication can become inactive. In such versatile environments, the latency to adapt to a time or space variation of the traffic distribution is an important parameter to consider.

The traffic produced by an application can be decomposed into elementary traffic types. Each elementary traffic type can be characterized by technical parameters determining system performance. They are:

- Throughput
- Message size
- User-level priority
- Time constraints: can be expressed by the maximum end-to-end delay, maximum tolerated jitter, or maximum time to establish or release a connection
- Dependability: defines the network's behavior when communication fails (message or connection loss, device failure, etc.)
- Type of communication: unicast, multicast, broadcast
- Message arrival law: periodic (with constant interarrival time), sporadic (with interarrival time greater than or equal to the minimum interarrival time), aperiodic (with only one arrival), bursty (with the duration and period of the burst), and so on

Wireless Standards and Protocols

We now introduce five wireless standards that are studied according to some constraints presented in the previous section. For global constraints, we focus on the topology used. For MAC-related constraints, the wireless standards are presented according to services provided for data transmission. In order to support traffic-related constraints, each wireless standard has to support the mapping functions of QoS requirements and mechanisms to support these QoS requirements.

IEEE 802.11

The IEEE 802.11 [1] protocol defines an ad hoc network architecture for devices within mutual communication range. When communicating devices belong to different coverage areas, distribution services are needed to communicate through a wired network. APs are used for bridging with wired networks and forming a WLAN.

IEEE 802.11 MAC can provide users two service types for data transmission: asynchronous transmission within a contention period, as provided by the distributed coordination function (DCF), and synchronous transmission within a contention-free period (CFP) as provided by the point coordination function (PCF). Integration of the two access methods is achieved using a superframe (Fig. 4), which includes two successive periods: a contention-free period and a contention one. In the situation illustrated in Fig. 4, the medium is busy when a new CFP starts. This period is therefore foreshortened.

For both services, the MAC protocol supports mechanisms such as the acknowledgment of unicast transmission and the handshaking control frames to reduce the hidden node problem. This mechanism, called virtual sensing, is based on the exchange of short control frames: a Request To Send (RTS) frame sent by a potential transmitter to the receiver and a Clear To Send (CTS) frame sent from the intended receiver in response to the received RTS frame. Any other node having received either this RTS or CTS frame delays its transmission. The effectiveness of the RTS/CTS mechanism depends on the length of the packet to be protected [2]. An RTS_threshold fixed by the user could be applied to disable the mechanism for shortest frames.

Moreover, different interframe spaces (IFS) have been introduced (Fig. 5): shortest for control frames such as acknowledgment and RTS/CTS frames, medium for synchronous traffic, and longest for asynchronous traffic. These IFS of different duration are a way to implement prioritized accesses. Lower-priority packets can be delayed with larger IFS [3].

The DCF implements the basic access method CSMA/CA. Collisions are avoided by means of a random backoff counter. The backoff period is bounded by two configurable parameters: the lower bound is Short Retry Limit (SRL) and the upper bound is Long Retry Limit (LRL), which control the latency and reliability of pending messages [4]. Moreover, the size of the backoff period and its offset can be made shorter to implement priority between different traffic types.

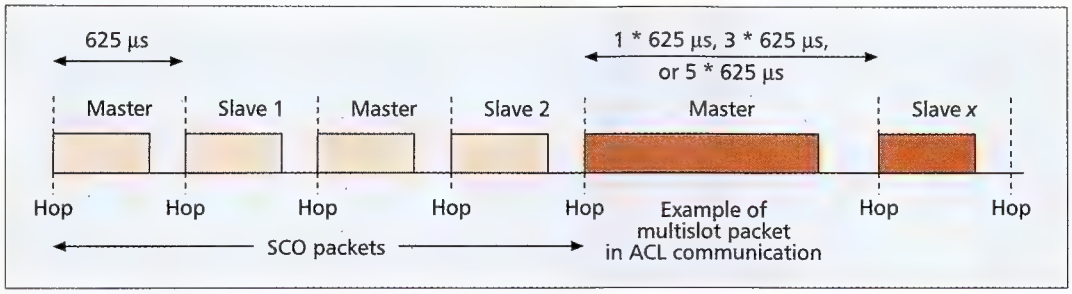


Figure 6. *TDD and timing in Bluetooth communication.*

Higher-priority packets will use a smaller backoff period, and the offset can be calculated such that there is no overlap in contention periods.

The PCF, which implements a polling access method, is available only in an infrastructured topology where an AP is used. The PCF relies on the service provided by the DCF, because devices requesting for synchronous communication have to use DCF rules to notify the AP. The AP periodically polls devices, giving them the opportunity to transmit. The maximum duration of CFP (CFP_Max_Duration) and its repetition rate (CFP_Rate), defined as a number of superframes, should be determined according to the characteristics of traffic that has to be conveyed by the AP. Other polling-based MAC mechanisms, more efficient than round-robin, can be implemented to support real-time communications (see, e.g., [5]). Proposals for a QoS enhanced PCF have been presented as part of the 802.11e Working Group.

Bluetooth

Bluetooth devices [6] use fast frequency hopping for data transmissions. Any two Bluetooth devices within range of each other can set up an ad hoc connection, establishing a small network called a piconet. Each piconet consists of a unique master that selects a frequency hopping sequence for the piconet and controls the access to the medium. Other participants of the piconet, called slave units, are synchronized to the hopping sequence of the piconet master. Up to seven active slave units may form a piconet. Multiple overlapping piconets can coexist because each piconet uses a different hopping sequence. Piconets can be interconnected via bridges to form a bigger network called a *scatternet*. However, to support forwarding data between devices from different piconets, a routing mechanism must be implemented at upper layers [7]. To deal with the hidden node problem, Bluetooth devices use fast frequency hopping.

To communicate, Bluetooth MAC implements a time-division duplex (TDD) scheme, splitting the channel into slots alternately used for transmission and reception between the piconet master and its slaves (Fig. 6). The slots for any master-to-slave communication are immediately followed by slots for slave-to-master communication. A slave is only allowed to transmit in a given slot if the master has polled it in a preceding slot. Moreover, any direct communication between slaves is impossible. The piconet master has to manage a polling scheme between slaves using a round-robin scheme or a more efficient scheme based on delay and throughput, as studied in [8].

Two types of service for data communications are provided to the users: asynchronous, called asynchronous connectionless (ACL), and synchronous, named synchronous connection-oriented (SCO). ACL service is used for the transmission of data bursts. The master entirely manages the ACL connection bandwidth: a master polls a slave, which may answer during following slot(s). The master defines the maximum packet length granted to the slave; this length is based on criteria such as the quantity of data to be transmitted or the number of successive slots available in the presence of SCO traffic. A maximum poll interval can be negotiated as the maximum time between subsequent transmissions between the master and a particular slave. It is used to support bandwidth allocation and latency control. The SCO service supports time constrained point-to-point connections. The length of an SCO packet cannot exceed one time slot. This service allows the negotiation of repetition rate and latency.

Best-effort QoS is generally applied to communications. A user can request a QoS guarantee for traffic defined by its flow spec parameters such as token rate and peak bandwidth. All these options have to be negotiated during connection establishment.

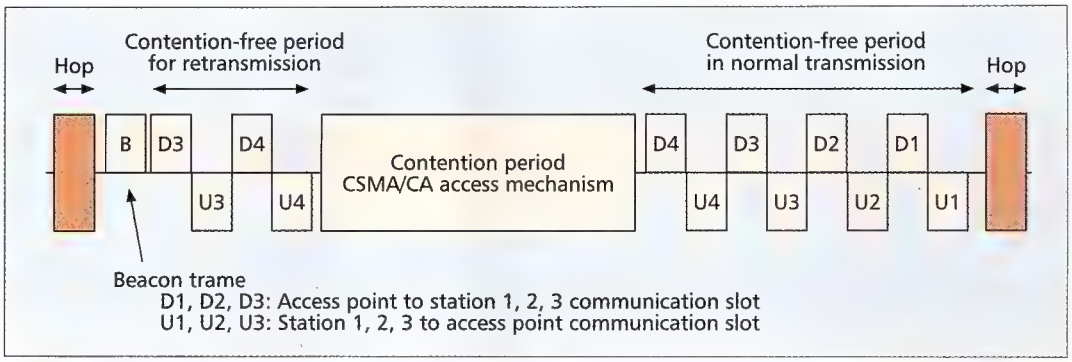
HomeRF

Based on the Shared Wireless Access Protocol (SWAP) 1.0 [9], HomeRF architecture is a combination of a managed network for synchronous and centralized services and an ad hoc network that provides asynchronous service. A unique AP, generally a gateway to a wired network, manages centralized services, and synchronizes and controls communications on the medium according to MAC rules. In a certain way, a unique AP prevents the hidden node problem. If the AP is not available, the devices may create an ad hoc network where the control of the network is distributed between all the devices.

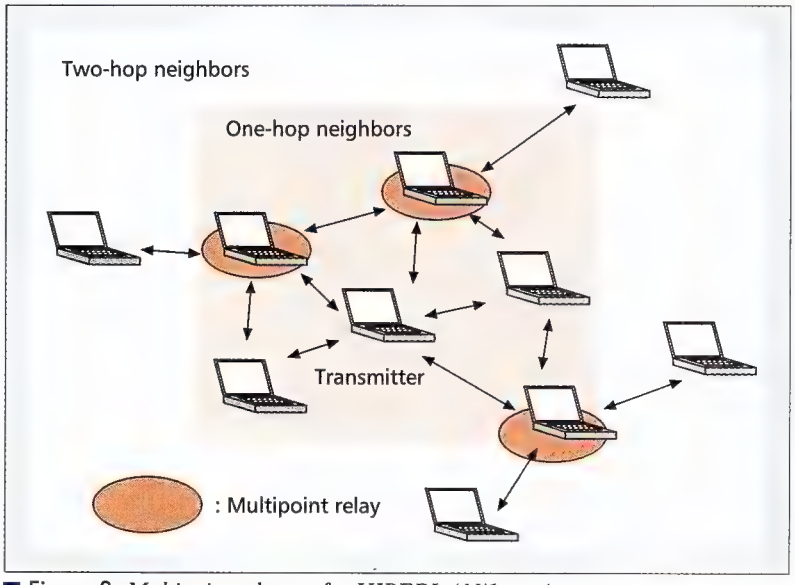
For communication services, a user may choose between two services types: synchronous and asynchronous. The SWAP MAC includes a time-division multiple access (TDMA) scheme to support synchronous traffic, as well as a CSMA/CA scheme derived from DCF in IEEE 802.11. This last scheme supports asynchronous traffic and control frames (e.g., request for synchronous traffic).

The MAC protocol uses a superframe incorporating two CFPs and one contention period (Fig. 7). The access mechanism used during each CFP is TDMA, while CSMA/CA is used during the contention period. The next version of HomeRF will implement a priority asynchronous data service in which selected data packets gain

Best-effort QoS is generally applied to communications. A user can request a QoS guarantee for traffic defined by its flow spec parameters such as token rate or peak bandwidth. All these options have to be negotiated during connection establishment.



■ **Figure 7.** *Basic MAC frame structure in HomeRF.*



■ Figure 8. *Multipoint relay set for HIPERLAN/1 routing.*

priority access to the channel by means of an offset on the contention period.

Each CFP is divided into six pairs of fixed-length slots. The first slot of each pair is used to transmit data from the AP to a device. Respectively, the second one is used for communications from a device to the AP. The second CFP, at the end of the superframe, is used for data transmission, while the first CFP at the start of the superframe is used for optional retransmission of any data not received or incorrectly received in the previous frame. To adapt the bandwidth to its requirements, a user can set the repetition rate of its allowed slots under the CFP mechanism.

HIPERLAN/1

The HIPERLAN/1 standard [10] works independent of any fixed infrastructure as an ad hoc network but allows access to conventional wired networks. HIPERLAN/1 supports the ad hoc topology with a multihop routing capability at the MAC layer. With this capacity, packets are forwarded from a source to a destination that cannot be reached directly. This routing protocol [11] is proactive: it maintains a route for each destination in the wireless network. It belongs to the link state family, in which the broadcast of topology information has been optimized. The multipoint relay set, a subset of one-hop neighbors of the transmitter, allows connection to all two-hop devices of the WLAN. This set has been introduced for that purpose: only the multipoint relays of the transmitter forward the packet, making this routing protocol efficient (Fig. 8).

The access scheme used in HIPERLAN/1 is called EY-NPMA, a specific implementation of the traditional CSMA with active signaling. This mechanism is a way to deal with the hidden node problem. With active signaling, a signal burst is sent by contending nodes before data transmission, and the usual carrier sensing is employed during idle time. The performance analysis given in [12] shows that active signaling has better performance than CSMA/CA with regard to the hidden node problem and unfairness in the medium access.

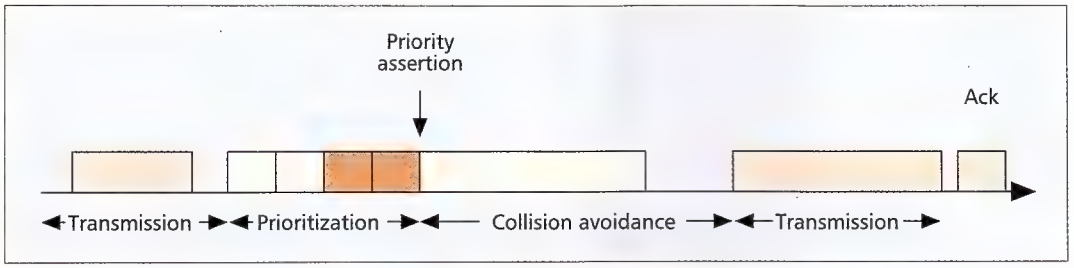
The MAC layer of HIPERLAN/1 uses a form of priority in granting access to the medium. Basically, when the channel is assessed to be clear, packets ready for transmission are submitted to a prioritization phase during which the packets with the highest priority are selected (Fig. 9). Packets with the same access priority are submitted to a collision avoidance phase. Each remaining device transmits a signal, the length of which is chosen in a random fashion. The device that broadcasts the longest signal gains control of the channel. Acknowledgment of unicast transmission is done after data transmission.

The two criteria used to establish the priority level of a packet are the user-level priority (as described in the traffic constraints discussion earlier) and residual packet lifetime (normalized to the number of hops to reach the final destination) deduced from the packet deadline given by the user. The mapping function of the user-level priority and normalized residual lifetime into the medium access priority is given in [8]. So HIPERLAN/1 offers prioritized access to fulfill traffic requirements for delay-sensitive applications.

HIPERLAN/2

The HiperLAN/2 standard [13] is mainly defined for infrastructure-based topology where an AP manages the wireless network. The AP controls in-range devices with connection-oriented links. This centralized topology is combined with an ad hoc capability: unlike some centralized protocols where two devices attached to the same AP cannot communicate directly, HIPERLAN/2 has an option allowing a direct communication between such in-range devices.

Several logical links can be established between devices and a priority, mapped from the user-level

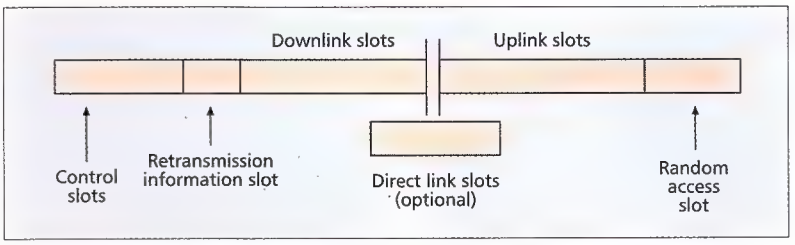


■ **Figure 9.** *The basic MAC frame structure in HIPERLAN/1.*

priority, is attached to each link. This priority can be translated into traffic classes according to IEEE 802.1p [14]. When a device wants to transmit data, it must first send a resource request to the AP in random access slots, which grants resources in the next frames according to the traffic, other requests, and its own transmission buffers. A request contains the number of packets to be sent, transmission frequency, and link quality requested. These parameters are determined according to the application's needs. The hidden node problem is solved by means of a centralized communication. Indeed, all communications are managed by the AP with a contention-free mechanism, even in direct communication between two devices.

The MAC protocol in the HIPERLAN/2 standard is based on TDMA. The basic frame structure has a fixed duration and incorporates several communications slots: one for broadcast and frame control, one for retransmission, two for data transmission (downlink and uplink directions), possibly one for direct mode transmission, and the last one for random access (Fig. 10). The AP can adapt the frame composition to traffic variation and adjust access delay. The scheduling policy, performed at the MAC level of the AP, is not specified. To support QoS requirements, additional mechanisms such as admission control and congestion control have to be implemented.

The connection-oriented nature of HIPERLAN/2 makes it straightforward to implement support for QoS. To adapt the quality and reliability of the link as a function of the environment, HIPERLAN/2 provides a set of internal functionalities such as dynamic frequency and data rate selection, and link adaptation. Depending on the performance and the reliability requested for transmission, the user can choose for each connection whether to use these previous mechanisms.



■ Figure 10. *The superframe structure in HIPERLAN/2.*

Adequacy between Studied Wireless Protocols and Application Constraints

Recapitulative Table

Table 1 shows how the studied wireless standards account for the previous application constraints: global, MAC-related, and traffic-related.

Adequacy Analysis

We summarize the main results described in Table 1.

If the application is such that all communicating devices are in range, any studied standard is possible. However, if the communication is not centralized (i.e., in the application any two devices can communicate directly), the Bluetooth and HIPERLAN/2 solutions without the direct communication option are less efficient. Indeed, with these two standards, any communication is not only controlled by the AP or the master but is received by this specific point and then resent to its destination.

On the other hand, if the application is such that out-of-range devices have to communicate and there is no infrastructure, routing is needed between all the devices of the WLAN. Routing can be implemented at either the MAC layer (e.g., HIPERLAN/1) or the upper layer (e.g., routing in the Mobile Ad Hoc Network, MANET, working group of the Internet Engineering Task Force, IETF). If an infrastructure exists, each AP is in charge of routing messages whose sender or destination does not belong to the devices it manages. This routing is achieved by a layer higher than two.

In the same way, mobility can be handled at layer 2 (e.g., HIPERLAN/1) or layer 3 (e.g., Mobile IP). As bandwidth is a killing resource in wireless networks, the routing traffic is reduced by limiting as much as possible the number of devices propagating the routing information.

Generally, the traffic induced by an application has a synchronous component and an asynchronous one. HIPERLAN/1 is the only wireless standard handling these two components in a uniform way: bandwidth is shared according to the message priority accounting for the message deadline and the user-level priority. All the other standards resort to two different MAC access schemes: polling for synchronous traffic and contention for asynchronous. When the synchronous traffic is well known and not subject to frequent changes, a polling scheme can guarantee low jitter and transmission delay for this traffic. However, it must be studied carefully in order to maximize bandwidth utilization: the slots allocated to polled devices with nothing to transmit must not be wasted. On the other hand, a MAC access scheme based on contention adapts itself immediately to traffic changes. However the contention scheme must be improved to manage different priorities,

enabling the message with the highest priority to be transmitted. The coexistence of both access schemes is made possible by the following rule: synchronous traffic is served first, whereas asynchronous traffic is served only if there is enough bandwidth left over by synchronous traffic. This rule is acceptable by an application only if this application has no real-time constraint attached to asynchronous traffic. Indeed, this traffic can suffer long queuing delays, increasing the jitter and end-to-end transmission delay.

Synchronous traffic can easily be differentiated from asynchronous traffic in almost all standards. Except for HIPERLAN/1 and 2, no mechanism

	IEEE 802.11	HomeRF	Bluetooth	HIPERLAN1	HIPERLAN2
Global constraints					
Nomadicity	Partially achieved by the AP	Only within the AP coverage	Partially achieved by the AP	Achieved by the MAC routing protocol	Partially achieved by the AP
Mobility	Yes, if managed by upper layers	No	Master mobility affects piconet configuration Must be managed by upper layer	Yes (for slow, medium speed)	Yes, if roaming by the fixed LAN
Power management	With awake patterns and AP buffers messages during sleeping period	With awake patterns and AP buffers messages during sleeping period	With awake patterns	With awake patterns and buffering devices	With awake patterns and AP buffers messages during sleeping period
Number of interconnected devices	127	127	7 slaves per master	–	256
Global throughput	1, 2 Mb/s for 802.11 11 Mb/s for 802.11b 54 Mb/s for 802.11g 54 Mb/s for 802.11a	1.6 Mb/s	1 Mb/s	23.5 Mb/s	54 Mb/s
Frequency band	5 GHz for 802.11a 2.4 GHz for the others	2.4 GHz	2.4 GHz	5 GHz	5 GHz
Confidentiality	Authentication + WEP	Authentication + encryption	Key sharing + encryption	Encryption	Key sharing + encryption
Easy installation	Immediate for ad hoc	Immediate for ad hoc	Yes	Immediate for ad hoc	Immediate for ad hoc
Easy maintenance	Yes	Yes	Yes	Yes	Yes
Cost effectiveness	Medium $150	Low-medium $10 to $100	Low $5 to $50	High	High
Ability to interconnect with other networks	With bridge for IEEE802 LAN, with router otherwise	With bridge for IEEE802 LAN, with router otherwise	With PPP protocols or with bridge for IEEE802 LAN	With bridge for IEEE802 LAN, with router otherwise	With internal functionalities to support packet- and cell-based traffic
MAC constraints					
Wireless topology	Ad hoc or infrastructure	Ad hoc or infrastructure with only one AP	Infrastructure	Ad hoc	Ad hoc or infrastructure
Medium access type for infrastructure topology	Decentralized for asynchronous traffic Centralized otherwise	Decentralized for asynchronous traffic Centralized otherwise	Centralized	Decentralized	Centralized
Medium access type for ad hoc topology	Decentralized for asynchronous traffic only	Decentralized for asynchronous traffic only	Centralized	Decentralized	Centralized with election of an AP for the control of the medium
Centralized communication	No	No	Yes	No	No if optional direct communication
Communication with out-of-range devices	Yes if upper layer routing is employed	Yes if upper layer routing is employed	Yes if upper layer routing is employed	Yes	Yes if upper layer routing is employed
Solution for hidden node problem	RTS and CTS frames	Only one AP used	Frequency hopping	Active signaling	Each neighbor uses a different channel frequency

Continued on next page

supports differentiation among asynchronous traffic. However, this differentiation is generally requested by multimedia applications. Additional mechanisms have to be implemented to support differentiation between asynchronous traffic.

HIPERLAN/2 and IEEE 802.11a are based on similar technology, operating in the 5 GHz band. HIPERLAN/2 has technical advantages, such as the inclusion of QoS mechanisms that allow it to handle voice and streaming media, and technology to prevent interference with other 5 GHz radio equipment; however, 802.11a arrived first on the U.S. market.

Open Issues

At the end of this study, we can point out some open issues. We can distinguish three major technical issues related to the MAC access scheme, routing and mobility, and the end-to-end QoS in multihop wireless networks:

The MAC access scheme and QoS requirements in a one-hop network:

Hidden node. How can we find an efficient solution to this problem, inherent to wireless networks?

Mapping of user priority and real-time constraints. The real-time constraints can be attached to synchronous and asynchronous traffic. In all studied standards but HIPERLAN/1, asynchronous traffic is transmitted only if synchronous traffic has left enough bandwidth. How can it be improved? How can we differentiate asynchronous traffic with different real-time constraints?

Expression of real-time guarantees. In all studied standards but HIPERLAN/1, we notice no real-time guarantees on delay and jitter for asynchronous traffic. How can we express bounded delays for such traffic? On the other hand, delay and bounded jitter are guaranteed only for already established synchronous traffic. Indeed, for submitting new synchronous traffic, an admission request, treated as asynchronous traffic, is used. How can we decrease the latency to establish new synchronous traffic?

Routing and mobility in multihop wireless networks:

Efficiency of routing done at the MAC level vs. at the IP level: Implementing routing at layer 2 spares processing time in layer 3. However, it supposes that routing algorithms are implemented in the network controller firmware. This is a more expensive solution than a layer 3 solution, generally implemented in software.

QoS routing in multihop wireless networks. This is needed in order to ensure end-to-end QoS in such networks. QoS routing must select a route that offers the features (e.g., bandwidth, delay) required by the requested QoS.

Mobility in multihop wireless networks. How can we guarantee that the QoS is ensured while

	IEEE 802.11	HomeRF	Bluetooth	HIPERLAN1	HIPERLAN2
Traffic					
Unicast Multicast Broadcast	Unicast Multicast Broadcast: between in-range devices	Unicast Multicast Broadcast: between in-range devices	Unicast: between a master and a slave Broadcast: master only	Unicast Multicast Broadcast	Multicast Broadcast: between an AP and devices Unicast: between any two devices in range of the AP
End-to-end delay guarantee	Between two in-range devices with polling	Between two in-range devices with polling	In a piconet only, with polling	Deadline-oriented but probabilistic access	Between two in-range devices with polling
Bounded jitter	Between two in-range devices with polling	Between two in-range devices with polling	Between two in-range devices with polling	Probabilistic	Between two in-range devices with polling
Loss control	With ARQ for unicast packets	With ARQ for unicast packets	With ARQ for unicast packets	With ARQ for unicast packets	With ARQ for unicast packets
User-level priority	No	No	No	Translated into a MAC access priority	Translated into a rank of connection establishment
Dynamics of traffic distribution	No delay for asynchronous traffic change Update of the polling table for synchronous traffic change	No delay for asynchronous traffic change Update of the polling table for synchronous traffic change	Update of the polling table for any traffic change	No latency	Update of the polling table for any traffic
Coexistence of different traffic types	With a superframe: polling for synchronous traffic, contention for asynchronous	With a superframe: polling for synchronous traffic, contention for asynchronous	Polling for synchronous and asynchronous traffics Jitter for asynchronous traffic function of maximum polling interval	Contention for both synchronous and asynchronous traffic	With a superframe: Polling for any traffic

Table 1. *Application constraints applied to studied wireless networks.*

In a hierarchical approach, more intelligence is put in the AP, less complexity in the other devices. As the number of APs is small with regard to the number of other devices, such an approach makes it possible to reduce the networking cost of the application.

the device is moving? Mobility and real-time constraints are difficult to reconcile.

End-to-end QoS in multihop wireless networks and traffic differentiation: How do we account for the distance a message has to cross to reach the final destination? Can it be mapped into a medium access priority? How do we extend QoS properties obtained locally to a multihop network? How can we keep determinism when an AP to a wired network is used? Furthermore, asynchronous traffic with real-time constraints or different priorities cannot be differentiated. How should we implement differentiation mechanisms in a current standard without drastic modification of the protocol stack?

From a more general point of view, two additional issues can be highlighted. They concern the design approaches in wireless networks and the compatibility of existing solutions:

Design approaches in wireless networks: In a hierarchical approach, more intelligence is put in the AP, less complexity in the other devices. Since the number of APs is small with regard to the number of other devices, such an approach makes it possible to reduce the networking cost of the application.

In a fully decentralized approach, all devices have the same intelligence. In order to reduce the amount of control traffic, they can elect a device among them to achieve particular centralized functions.

Compatibility of different solutions: A problem to be discussed concerns the compatibility of different products. It can be discussed at two levels:

- Coexistence and interference of different wireless networks in the same environment
- Interconnection of different heterogeneous networks

Conclusion and Perspectives

Wireless networks are now easy to deploy, and the available bandwidth becomes acceptable for most applications. Moreover, with wireless networks, mobility has become a reality. At the same time, the need for wireless products is increasing, based on different wireless protocols. Standardization is evolving quickly (e.g., HomeRF v. 2.0, Bluetooth as discussed in the IEEE 802.15 Personal Area Network Working Group, IEEE 802.11a). This article provides a comparison of these wireless protocols. We have focused on the first two layers of wireless standards: IEEE 802.11, HomeRF, Bluetooth, and HIPERLAN/1 and 2.

An important open issue deals with the coexistence of different traffic types. For instance, all studied protocols except HIPERLAN/1 favor synchronous traffic over asynchronous traffic, which is transmitted only if there is sufficient bandwidth left over by synchronous traffic. In order to support multimedia applications, service differentiation and traffic classes have to be provided by the underlying wireless network. To achieve that, priority-based medium access and efficient management of the waiting queues are needed. Some other technical issues require further study, concerning routing and mobility in multihop wireless networks. Furthermore, the design approaches in wireless networks and the compatibility of different existing solutions are still open issues.

References

[1] IEEE 802.11, "LAN MAN Standards Committee of the IEEE Computer Society. Wireless LAN Medium Access Control and Physical Layer Specifications," 1997.

[2] Y. Wang and B. Bensaou, "Priority Based Multiple Access for Service Differentiation in Wireless Ad-Hoc Networks," *Networking 2000*, Paris, France, 2000.

[3] J. Weinmiller *et al.*, "Analyzing the RTS/CTS Mechanism in the DFWMAC Media Access Protocol for Wireless LANs," *IFIP TC6 Wksp. Pers. WL Commun., Prague*, Czech Rep., 1995.

[4] I. Aad and C. Castelluccia, "Introducing Service Differentiation into IEEE 802.11," *Proc. 5th IEEE Symp. Comp. and Commun.*, Antibes, France, 2000.

[5] A. Köpsel, J. P. Ebert, and A. Wolisz, "A Performance Comparison of Point and Distributed Coordination Function of an IEEE 802.11 WLAN in the Presence of Real-Time Requirements," *Proc. 7th Int'l. Wksp. Mobile Multimedia Commun.*, Japan, 2000.

[6] Bluetooth SIG, "Specifications of the Bluetooth System 1.0b, vol. 1: Core," http://www.bluetooth.com, 1999.

[7] T. Salonidis *et al.*, "Distributed Topology Construction of Bluetooth Personal Area Networks," *IEEE INFOCOM*, Anchorage, AK, 2001.

[8] A.Capone, M. Gerla, R. Kapoor, "Efficient Polling Schemes for Bluetooth Piconets," *IEEE ICC '01*, Helsinki, Finland, 2001.

[9] Z. J. Negus, A. Stephens, and J. Lansford, "HomeRF and SWAP: Wireless Networking for the Connected Home," *ACM SIGMOBILE Mobile Comp. and Commun. Rev.*, 1998, vol. 2, pp. 28–37.

[10] ETSI STS-RES10 Tech. Comm., "Radio Equipment and Systems: High Performance Radio Local Area Network Type 1, Functional Specifications," 1996.

[11] P, Jacquet *et al.*, "Increasing Reliability in Cable-FreeRadio LANs, Low Level Forwarding in HIPERLAN" and "Data Transfer for HIPERLAN," *WL Pers. Commun.*, vol. 4, no. 1, 1997, pp. 51–80.

[12] A. Qayyum, "Analysis and Evaluation of Channel Access Schemes and Routing Protocols in Wireless LANs," Ph.D. thesis, Univ. of Paris-Sud, 2000.

[13] ETSI TS-101 761-1, "Technical Specification Broadband Radio Access Networks, HIPERLAN Type 2. Data Link Control Layer Part 1: Basic Data Transport Functions," 2000.

[14] IEEE 802.1D, "IEEE Standard for Local Area Network MAC Bridges," 1998.

Biographies

ANTOINE MERCIER (mercier@ece.fr) graduated as an engineer of ENSEA in 2000 from the Computer and Telecommunication Sciences School, France. He is preparing a Ph.D. of sciences on quality of service in wireless network protocols at the Univ. of Paris 12, France. He is now teaching as an assistant professor of network communication at ECE, Computer and Telecommunication Sciences School, France.

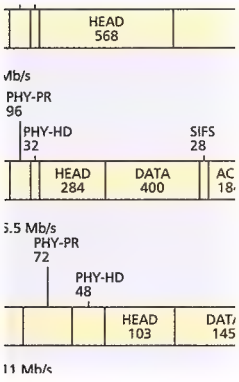
PASCALE MINET (pascale.minet@inria.fr) graduated with a Ph.D. in 1982 from the Univ. of Toulouse and as an engineer in computer science in 1980. She is a senior researcher with project Reflecs at INRIA, Rocquencourt, France. Her major research area is fault tolerance, real-time communication protocols, and concurrency control in distributed real-time systems. She is interested in the design methodology of real-time applications. She has been a member of the ETSI RES10 committee for the HIPERLAN standard which specifies a high-speed wireless LAN.

LAURENT GEORGE (george@univ-paris12.fr) prepared his Ph.D. thesis at INRIA. He obtained his thesis in January 1998 from the University of Versailles. The subjects of his thesis were online distributed real-time scheduling with consistency constraints and reliable/atomic multicast with real-time constraints. He is now teaching network communication at the University of Paris 12 as an assistant professor. His research activity concerns QoS for real-time distributed systems. He studies the problem of multimedia communication (over the Internet) and wireless communication with end-to-end real-time constraints.

GILLES MERCIER (mercier@univ-paris12.fr) has a 23-year career in computer and engineering systems, and since 1995 has been a professor at the University of Paris 12, mainly in sciences and computers, and real-time processes. He is a graduate engineer in electronics since 1978 and has a Ph.D. of sciences in real-time neural networks. His research topics focus on real-time industrial processes based on short distance wireless LAN communications for embedded systems.

VERTICAL OPTIMIZATION OF DATA TRANSMISSION FOR MOBILE WIRELESS TERMINALS

MICHAEL METHFESSEL, KAI F. DOMBROWSKI, PETER LANGENDÖRFER, HORST FRANKENFELDT, IRINA BABANSKAJA, IRINA MATTHAEI, AND ROLF KRAEMER, IHP



In addition to the convenient access to Internet applications from mobile devices, new types of service are expected to arise with a wide range of requirements on the mobile terminal. A major bottleneck is the need to develop handheld devices with enough processing power and battery life to support these applications.

ABSTRACT

A major problem for TCP connections over wireless links is that errors introduced by the wireless channel interfere with the TCP protocol, leading to reduced data rates and power wastage. Based on accurate simulations for the TCP and IEEE 802.11 MAC protocols, we discuss recipes to optimize transmission. It is argued that the best approach is to restrict modifications to the mobile device. While this requires separate solutions for the uplink and downlink, the results of optimization are then available when roaming into any WLAN obeying the relevant MAC protocol. Simulation results show that the combination of specific strategies with a vertical interaction between the protocol layers can lead to the required improvements, giving a promising approach to enhance the performance of wireless mobile terminals.

INTRODUCTION

The integration of wireless and mobile systems with the existing Internet requires small power-efficient easy-to-use mobile terminals with high processing power. In addition to convenient access to today's Internet applications from mobile devices, new types of service are expected to arise with a wide range of requirements on the mobile terminal. A major bottleneck is the need to develop handheld devices with enough processing power and battery life to support these applications. Especially at higher data rates, a large part of the resources will be taken up by communication over the wireless channel. Therefore, it is advisable to reduce this burden as much as possible when implementing the relevant protocols.

In this article the focus is on mobile devices used within a wireless local area network (WLAN) to access the Internet, whereby TCP is used as the end-to-end transport protocol between the device and the remote server. Data transfer therefore runs through a wireless channel followed by a fixed network Internet connection. Whereas the wireless hop is characterized by time-varying and high packet losses, the Internet connection suffers from unpredictable delays and possible congestion collapses. In this situation, it is well known that the losses on the wireless link interfere with the TCP protocol and can lead to drastic reduction of the data rate. Thus, the aim is to make the wireless link as efficient as possible by itself while at the same time minimizing the negative effect of wireless losses on the TCP protocol.

Numerous approaches have been suggested to cope with the situation, involving modifications to either the TCP protocol [1, 2] or the link layer on the mobile device and the base station [3–7]. This article discusses a more conservative approach in which only the mobile device is available to be modified. While this restricts the possibilities to some extent, it has the advantage that the benefits of optimization are available in all WLANs without introducing any new features to the protocols. The price is that two separate tasks must be solved: optimization of the wireless communication in the uplink and in the downlink.

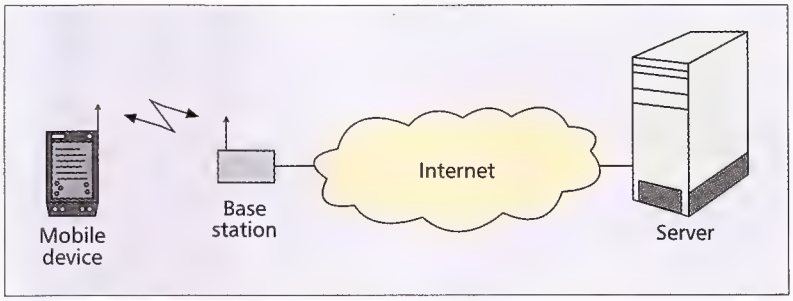
For the uplink, the mobile device has control over the transmission parameters. This case is best handled by building on previous work concerning adaptive link layer strategies [4, 5], adjusted to the specific MAC protocol used. For the downlink, new complementary strategies must be developed. In both cases, the protocol layers on the mobile device are allowed to interact, which leads to substantial performance improvements. At the lowest level, this vertical protocol optimization allows the lower layers to be aware of the type of data and specific quality of service (QoS) requirements requested by the application. At a higher level, vertical interaction permits some completely new approaches with significant benefits.

The results presented here were obtained using simulations that combine a working TCP implementation with an accurate model of the IEEE 802.11 MAC layer [8]. They show that the combination of specific strategies for the uplink with vertical optimization across the protocol stack on the mobile device can lead to great

1536-1284/02/$17.00 © 2002 IEEE

improvement in the performance and power efficiency of a mobile terminal. Some of the material included here has been presented previously [9].

The rest of this article is organized as follows. First, the approach outlined above is discussed in more detail, and the interference between the errors on the wireless channel and TCP for the specific system of interest is addressed. Next, simulation results are presented in the context of uplink optimization for the specific case of an IEEE 802.11 and 802.11b MAC protocol, deferring the results for the downlink to a future publication. Finally, the main results are summarized.



■ Figure 1. *A sketch of the network architecture considered in this article. The end-to-end TCP connection between a mobile device and a remote server runs through a wireless link in a local WLAN and the Internet.*

Protocol Optimization on a Mobile End Device

Figure 1 shows a typical scenario involving a mobile end device. From the user's viewpoint, the power consumption of the mobile device is to be reduced and overall performance optimized in terms of parameters such as availability, throughput, delay, and jitter. The mobile device is used to enter a LAN via a wireless connection to a base station. From there, the connection continues through the Internet to a server at some remote location. TCP runs as an end-to-end protocol on the mobile device and server, while a wireless protocol such as IEEE 802.11 controls transmission over the local wireless link. The wireless hop is prone to transmission errors from effects such as multipath propagation, shadowing, and interference from other systems sharing the same frequency band. Data transmission through the Internet can involve longer and unpredictable delays, and can collapse when congestion occurs. The combination of the two links with different characteristics, with TCP running end to end over both, makes the problem interesting.

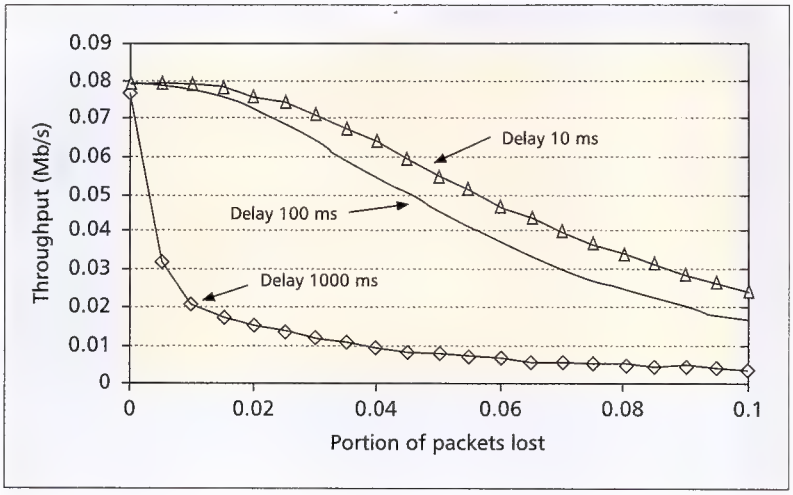
Two issues are discussed next in this context: general strategies for optimizing the performance, and the well-known problem of interference between wireless channel errors and the TCP protocol [10].

Optimization Strategies for the Mobile Device

A first point to note is that it could be useful to optimize performance by modifying the end device only. That is, it is assumed that the base station and remote server obey the relevant wireless MAC and TCP protocols, respectively, but otherwise are not available for modification. While this restricts the possibilities, it could be considerably more useful in practice. A main part of the incentive for mobile devices is the ability to roam through different networks. Thus, reliance on nonstandard extensions to the base station should be avoided. By the same logic, performance enhancements that require extensions to the TCP protocol will be difficult to apply in practice. On the other hand, the mobile device is completely available for modification as long as it obeys the MAC and TCP/IP standards itself. Specifically, there is freedom to let the layers interact vertically in the protocol stack on the mobile device if this turns out to be useful.

As a consequence of restricting the optimization to the mobile device, there are now two distinct problems to solve: the optimization of the wireless hop for the uplink and that for the downlink. For the most part, solutions applicable to one of these two cases will not be applicable to the other. For example, it is possible to fragmentize packets or to switch to a different error correction scheme in response to bad channel conditions when transmitting data from the mobile. However, if control is only over the mobile device, it cannot be assumed that the base station will do the same in the downlink. Instead, the base station will choose its own transmission characteristics by an algorithm that may or may not take channel conditions into account. Depending on the type of application, it could be more important to improve performance for either the uplink or downlink; for example, in scenarios such as Web browsing the downlink plays a larger role. Assuming symmetry in the data flow on average (also in view of future applications that go beyond Web browsing), it is concluded that it will be necessary to develop novel approaches for both directions in order to obtain an overall gain that is significantly larger than 50 percent.

In general terms, strategies to improve transmission and save power in the uplink work on the link layer by sensing the channel quality and adjusting the transmission parameters accordingly. Specific possibilities (see [11] for a survey) are use of a more error-resistant modulation scheme, increased forward error correction (FEC), retransmission of lost packets, division of packets into smaller fragments [4], and channel probing [12]. One challenge here is to implement an "intelligent" link layer that dynamically chooses the optimal behavior as channel quality changes. In the downlink, a major issue is the need to monitor the channel continuously for incoming data. This can add up to a large part of the total power consumption, even if no data is being received or sent by the mobile device in question. One approach to this problem is extensive use of sleep or doze modes. A second downlink issue is to increase the possibility of error-free reception of a data packet under high bit error rates, by "trying harder" in the receiver. In this article we address only the question of optimizing the uplink. Work is in progress on developing downlink strategies and will be presented in a separate article.



■ Figure 2. *Response of TCP throughput to random loss of packets in the uplink of the wireless hop for the system in Fig. 1 for a network delay of 1000, 100, and 10 ms, respectively. Each value shown is the average of 10 simulations for the transfer of 2 Mbytes. The TCP packet length was 500 bytes.*

Vertical protocol optimization [13] addresses the question in two different ways. First, the methods described in the previous paragraph involve a tradeoff between different QoS parameters such as data rate, delay, jitter, and possible loss or distortion of data. Depending on the type of application, restrictions are posed on the available range of the QoS parameters. For example, transmission of an email message can be delayed by a much larger time than streaming data, but cannot tolerate any distortion or loss. Thus, a vertical interaction between the application and the lower protocol layers is required to inform something like an "intelligent link layer" of the desired QoS. Second, by allowing the layers to interact, new possibilities for optimization arise that can lead to substantial benefit.

Interference between TCP and the Wireless Link

As a second point in this general discussion, the degradation of TCP throughput under wireless losses is considered for the specific arrangement shown in Fig. 1. The basic problem has been discussed many times and is an active area of research: the throughput can be decreased drastically even if only a few packets are lost in the channel. The reason is that the TCP protocol cannot distinguish packets lost somewhere in the Internet from those lost in the wireless link. Consequently, losses in the wireless transmission are incorrectly interpreted as a sign of congestion, leading to a reduction of the data flow by the TCP congestion control mechanism. Technically, the procedure is that the congestion window as well as the threshold that governs its rate of growth are reduced at each transmission error, for both retransmission timeouts and fast retransmissions. If this happens frequently, the congestion window becomes smaller than the capacity of the pipe, given by twice the delay-bandwidth product. From that point on the pipe is not kept full, and only a part of the available bandwidth is utilized.

For the system in Fig. 1, the focus of this discussion is on the total delay introduced by the Internet part. Since this delay enters directly into the calculation of the pipe capacity, it will have an influence on the TCP throughput problem. The aim is to see how severe the problem actually is for delays corresponding to modern Internet characteristics. The simulation results in Fig. 2 show how the throughput decreases as packets are lost in the wireless link for three different values of the network delay. For a delay of 1000 ms, there is the expected extreme sensitivity to packet loss. For example, the loss of one packet in 300 already reduces the data rate to one-third. However, it turns out that the effect is considerably less severe when the network delay is 100 or 10 ms, as can be seen from the top two curves in Fig. 2. As discussed in more detail in [14], the reason for the different behavior is as follows. For the large network delay of 1000 ms, the pipe capacity is 20 kbytes. Repeated transmission errors keep the congestion window down to a few kilobytes, so only a small part of the available bandwidth is actually used. For delays of 100 and 10 ms, the pipe capacity is only 2.0 and 0.2 kbytes, respectively. The congestion window is still kept down to a few kilobytes on average by the transmission errors, but this does not restrict the data rate as in the previous case. Instead, the comparatively mild reduction of the throughput seen here is a combination of various effects. First, each fast retransmit operation leads to some extra delay in addition to that needed to actually retransmit the relevant packet when the send window is exhausted. Second, for certain patterns in packet loss, the fast retransmit mechanism is not able to cope, and data transfer recommences only after a retransmission timeout (RTO).

These results suggest that the well-known problem of a wireless link interfering with the TCP protocol could be less severe than expected in some situations of practical relevance. First, in a scenario whereby a mobile end device is used to access the Internet as in Fig. 1, the TCP performance could already be adequate if the network delay is reasonably small, even if the wireless hop has a relatively high bit error rate. Second, if TCP is used locally within a wireless LAN, delays are so short that there will be even less of a problem. Of course, the validity of these conclusions also depends on specifics such as data rate, buffer sizes, and the flavor of TCP used. However, it seems advisable to inspect the issue carefully for a given system of interest.

Simulation Setup: TCP Running over the IEEE 802.11 MAC

In the following, results are presented for simulations that combine a realistic TCP implementation with a model of the IEEE 802.11 and 11b MAC layer. The aim of the study is to help develop specific optimization strategies for the uplink from a mobile device, for a system of practical relevance. The focus is on the retry effort on the MAC layer and on fragmentation as a means to reduce packet loss in the wireless link. Although such questions have been consid-

ered previously, our results are valuable because they are based on accurate implementations of the relevant protocols.

The simulations were done using the BONeS tool from Cadence. A C implementation of TCP in the New Reno variant was used. This implementation was developed specifically for handheld devices with restricted resources. It has been checked by extensive interoperability tests against Linux, FreeBSD, and Microsoft versions of TCP under a lossy channel. The MAC layer was implemented in the BONeS language, based on a model of the PHY layer that takes account the correct timings and delays. The IP and LLC layers were included in a rudimentary form only since they do not significantly influence the throughput.

The simulated system corresponds to that shown in Fig. 1. For the questions of interest here, some components in the three participating devices (mobile, base station, and server) could be discarded to avoid simulation overhead. These include half of the base station (which communicates with the Internet) and the lower protocol layers of the server. More precisely, only two complete protocol stacks were used, but with an "Internet channel" block inserted between the LLC and IP layers of the second. This stack therefore combines the lower layers of the base station with the upper layers of the server. The Internet channel is characterized by two parameters: the network data rate and the network delay. It contains a queue in each direction that is emptied packet by packet as determined by packet length and network data rate. Each packet is then additionally delayed by the network delay before it is passed on to the next block. The single queue in each direction models all router queues encountered between the base station and the server.

As discussed above, the problem of interference between TCP and the wireless channel is most significant if the pipe is reasonably large. Therefore, the simulation parameters were chosen to model this case. The network delay was kept at 1000 ms (except for the simulations at 100 and 10 ms described in the previous section and where otherwise stated), and the network data rate was set to 0.08 Mb/s, yielding a pipe capacity of 20 kbytes. This value was also selected for the size of the TCP send and receive buffers. The data rate of the wireless transmission was set to 2 Mb/s. This parameter does not play a very large role since the overall rate is restricted by the much lower network data rate. Not quite as obviously, the network data rate itself can be increased without significantly changing the results shown below. The reason is that the rate by which TCP transmits is limited to the maximum amount of data allowed in flight, divided by twice the network delay. If the in-flight data is limited by the buffer size instead of the pipe capacity, basically the same overall behavior results. The main difference is that data is sent in separated bursts instead of a continuous stream.

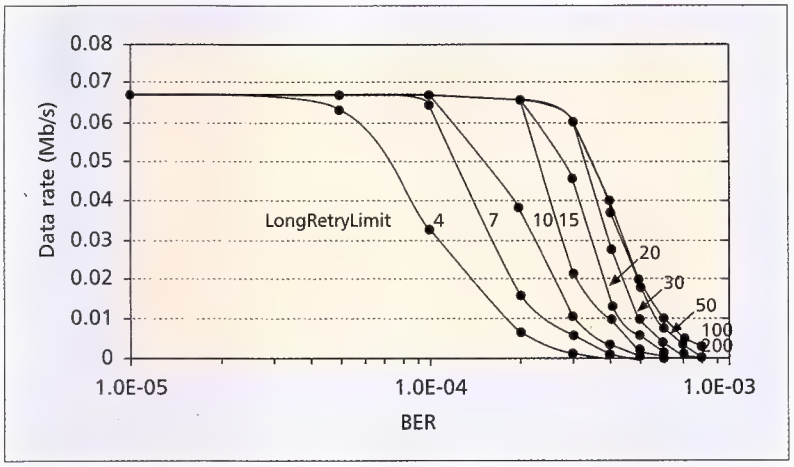


Figure 3. *TCP throughput as a function of the BER in the wireless channel for various values of the MAC long retry limit. The results were obtained using simulations for the system shown in Fig. 1 as described in the text.*

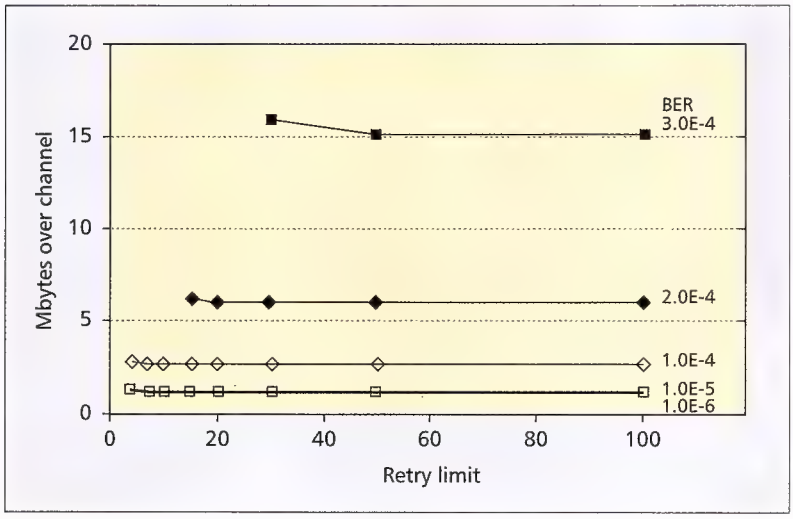
Optimizing the Uplink

Considerable earlier work has been done on adaptive link layer strategies [3–5] as a means to improve transmission over a time-varying faulty wireless channel. By sensing the quality of the channel and adjusting the transmission characteristics accordingly, power can be saved and data transfer improved. Specific mechanisms include packet fragmentation, switching to a slower but more robust data rate, modifying the error correction mechanism, and channel probing to detect the end of a channel downtime efficiently. These approaches are generally introduced in both directions, that is, on the base station as well as on the mobile device. Consistent with the philosophy of this article, they are considered here only as a means to improve the uplink from the mobile to the base station.

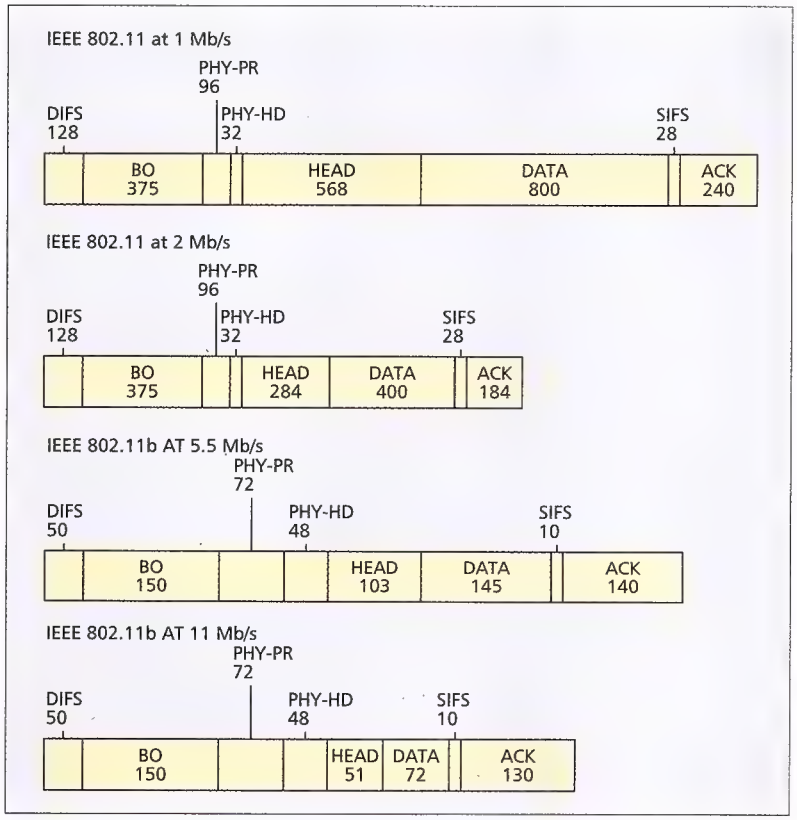
A dynamic adaptation of the transmission characteristics inherently involves complicated trade-offs. One can increase the error correction only at the price of a lower data rate, and fragment the packets only at the price of increased overhead. Since different mechanisms are available, one difficulty is to choose among these in a certain situation. For example, assume that the channel starts off with high quality, but then deteriorates to lower quality. By switching to a lower data rate and retransmitting packets continuously, the transmitter can ensure that data still arrives. However, the cost is additional power consumption, since the RF transmitter is turned on for a longer time per packet and energy is wasted in each unsuccessful attempt. Thus, it could be better to stop transmitting completely (except for spaced-out probe packets) until a substantial improvement in channel quality is sensed. Unfortunately, this could happen at some indefinite time in the future. Thus, a decision must be made based on at least two thresholds: how bad the channel must be before switching to a different method and the maximum permitted transmission delay. As a final complication, better performance might be reached if a history of channel behavior is maintained.

An effective adaptive link layer therefore involves something close to an expert system that makes complicated decisions based on multidimensional input data. As pointed out in

[15], awareness of the QoS requirements for the relevant data stream should be included. This information defines the limits when adjusting the characteristics of the transmission (e.g., the lowest data rate that is still useful, maximum permitted delay, and whether some data loss is allowed). Since the QoS parameters are defined by the application, this is a straightforward example of vertical interaction between the protocol layers. With such a system as the final goal, the first step is to collect information on the improvement attainable when no such restrictions are taken into account. The results presented in the rest of the section contribute to this in the context of link layer retransmission and fragmentation. Within the vertical optimization framework, it is assumed that the optimization is specifically for TCP, ignoring any disadvantages arising for other protocols such as UDP.



■ Figure 4. *Total number of bytes transmitted into the wireless channel as a measure of the consumed energy for the same simulations as in Fig. 3. The results show that for a given BER, the consumed energy is independent of the MAC long retry limit.*



■ Figure 5. *Required time in microseconds at the PHY level for the various phases in the transmission of a packet with 100 bytes of payload for the IEEE 802.11 standard at 1 and 2 Mb/s (top) and the 802.11b standard at 5.5 and 11 Mb/s (bottom). From left to right: MAC DIFS idle period, average backoff time, PHY preamble and header, sum of TCP/IP/DL/MAC headers, user data, MAC SIFS delay, and total time used for the MAC acknowledgment. Note that different horizontal scales are used for the 802.11 and 802.11b cases.*

Retransmission on the TCP and MAC Layers

When a packet is lost on the wireless link (by either a collision or channel error) the MAC layer retransmits the packet up to a maximum number of attempts (called aLongRetryLimit in the standard [8] and LRL here). When the limit is reached, the packet is discarded and the procedure restarted with the next packet in the output queue. However, a central task of the TCP protocol is to ensure that each packet arrives sooner or later. Thus, when the MAC layer discards a packet, the consequence is that TCP must retransmit it, albeit at a later time. In other words, every packet must be transmitted repeatedly until it finally gets through, and the only question is which part of the protocol stack does this.

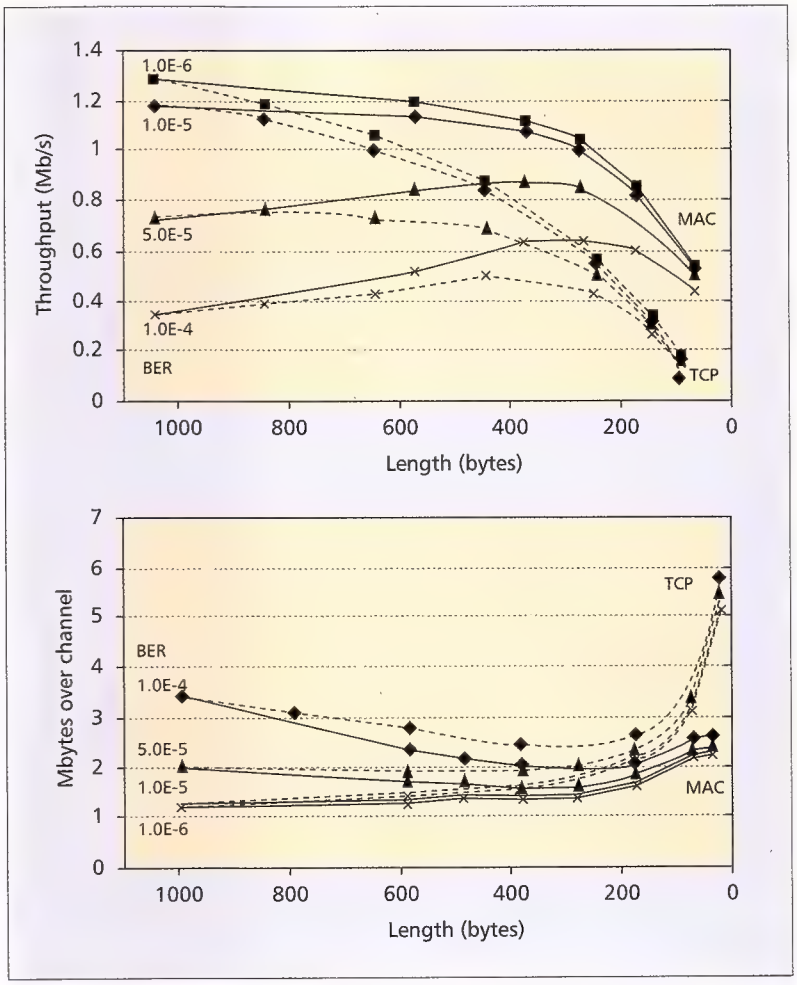
By varying the LRL in the simulation, the burden of retransmitting lost packets can be shifted between the MAC and TCP layers. The results are displayed in two ways for simulations based on the 802.11 standard. Due to the low nominal network data rate of 0.08 Mb/s considered here (see above), a faster wireless link does not change the situation, so the results for the 802.11b standard are almost identical. Figure 3 shows the throughput (i.e., the actually attained data rate) as function of the bit error rate (BER) for various choices of the LRL. Evidently the value of LRL makes a significant difference in the throughput, whereby the best performance results when LRL is large. In this limit the MAC layer patiently continues to retransmit a given packet even if a very large number of attempts are needed in the end. Figure 4 shows the total number of bytes sent into the wireless channel, which is a measure of the energy consumed during transmission. In contrast to the data rate, these curves turns out to be almost flat (i.e., independent of the LRL). Thus, for a given packet length, coding scheme, and BER, the consumed energy is a fixed quantity, while the required transmission time becomes significantly larger if TCP takes care of retransmissions.

The flat curves look somewhat surprising at first, considering that the system really is behaving in a different way when the LRL is changed. However, the result is easy to understand by noting that the average number of attempts needed to get through is only a property of the wireless channel, and not of the transmitting system. More precisely, at a given BER there is

some fixed probability P that a packet is corrupted during the transmission. The average number of attempts needed to get through is then $1/(1 - P)$. From this it follows that the throughput for the simplest possible transport protocol (retransmission of each packet without any upper limit until it is finally received correctly) equals the nominal rate of the wireless channel $\times 1 - P$. When this expression is truncated vertically at the network data rate used in the simulation, it closely reproduces the limiting curve for large LRL in Fig. 3.

The conclusion is that there will always be the loss of a certain amount of energy as the packet is repeatedly sent into the faulty channel until it finally gets through, but the situation is always made worse if this is done by TCP instead of the MAC layer. It follows that it is pointless to ever discard a TCP packet in the MAC layer. Every discarded packet will only be retransmitted by TCP later on, but it will then additionally interfere with the internals of the TCP protocol by triggering a fast retransmit or a retransmission timeout event. It is the delays associated with these events that cause the reduced data rate when LRL is smaller. The optimal MAC strategy for TCP data streams could thus be called "never give up": there is no point in ever dropping a TCP packet and proceeding to the next one on the MAC level. Of course, this only becomes an issue if the channel is so faulty that the standard MAC retransmission strategy (typically 8 attempts/packet) is still not adequate to ensure correct reception of the packets.

The "never give up" argument is of more general validity beyond the question of the retry limit. When a TCP packet has difficulties in getting through the wireless channel, the MAC layer should use all means at its disposal to ensure correct delivery of this specific packet. This could involve extreme fragmentation, low data rates, increased forward error correction, or all of these at the same time. Timeouts at the TCP level are generally much larger than the delays in the wireless channel, so there is enough time available for numerous attempts. This basic strategy, while simple in principle, requires additional information from other protocol layers within the MAC layer. First of all, a packet must be identifiable as part of a TCP stream. Second, the QoS limitations should be known since these could be exceeded during the retransmission effort. Finally, as discussed above, the reduction of the TCP data rate is not too extreme in some cases; for these, it might be better to live with the reduction instead of making the additional effort at the MAC level.

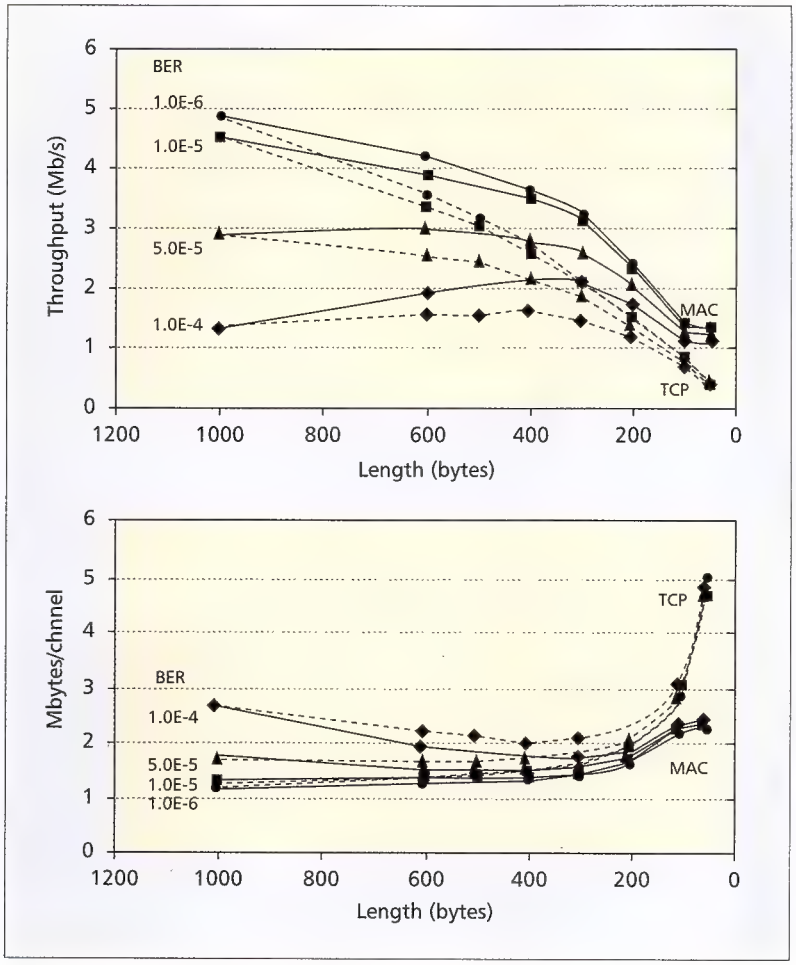


■ Figure 6. *Top: comparison of the TCP throughput for the case of fragmentation on the TCP layer (dashed lines) and on the MAC layer (continuous lines) as a function of the payload per packet for various wireless channel BERs in the IEEE 802.11 standard. Bottom: the total bytes transmitted into the wireless channel for the same simulations as a measure of the consumed energy.*

Packet Fragmentation

In the case of good channel conditions, the only effect of reducing the packet length is to increase the overhead needed to transmit a certain chunk of data. But if the channel becomes bad, large packets are more likely to be damaged, increasing the number of attempts needed to transmit the packets correctly. This slows down transmission and increases power consumption. In order to minimize these effects it makes sense to adapt the packet length to the channel conditions. This involves a trade-off between the improved probability that the packet is transmitted correctly and the increased overhead. One question of interest is whether a reasonable compromise can be found that gives adequate performance over a wide range of BERs, or whether a significant further gain can be obtained by choosing the length dynamically based on error rate. The question of the trade-off when fragmenting packets has been addressed before [4]. The value of the investigations presented here is that they were obtained for TCP running over an accurate model for a popular MAC standard.

The packet length can be changed by either reducing the size of a TCP packet or introducing fragmentation at the MAC level. The most important difference is the amount of overhead added to the total data amount by each newly created packet. For example, for the IEEE 802.11 standard at 2 Mb/s, Fig. 5 shows that transmission of the overhead takes 1127 µs, while the user data requires 4000 and 400 µs for payloads of 1000 and 100 bytes, respectively. The percentage used to transmit actual user data thus reduces from 78 percent to only 8 percent.



■ Figure 7. *Comparison of the effect of fragmentation on the TCP and MAC layer on the throughput (top) and number of bytes transmitted into the channel (bottom) as in Fig. 6, but for the IEEE 802.11b standard at the nominal data rate 11 Mb/s.*

The point is that a reasonable degree of fragmentation quickly leads to a situation in which the user data portion is small compared to the overhead. The lower half of Fig. 6 shows this data for the case of the IEEE 802.11b standard. Although there are some differences in the way in which the total required transmission time splits up, the overall picture is similar. In either case, the overhead added by fragmentation on the MAC level is smaller than for TCP since it avoids the higher-level headers, SIFS delay, and backoff slots.

The effect of TCP and MAC fragmentation on data throughput at various error rates is compared in Fig. 6 for the 802.11 standard. In these simulations, the network delay was reduced to 1 ms. The overall throughput is then governed by the wireless link, not by the amount of TCP in-flight data. Performance is always better for MAC fragmentation as expected. While the overhead is acceptable down to payloads of about 200 bytes for the MAC case, TCP fragmentation throughput decreases steadily with reduced packet size. Thus, fragmentation at the MAC layer is clearly preferable, leading to an improvement by a factor of two to three in some cases. Furthermore, the curves show that a fixed payload length of around 300 bytes gives a reasonable performance over a wide range of channel conditions when MAC fragmentation is used, whereas a fixed choice is difficult to find for TCP layer fragmentation. The lower half of Fig. 6 shows that basically the same conclusions apply when the focus is on energy consumption instead of throughput. Finally, equivalent simulations were done for the 802.11b standard, yielding the results displayed in Fig. 7. Except for the overall higher data rates, the pictures look very similar, so the same conclusions apply to this standard.

Summary and Conclusions

The basic idea of the approach discussed in this article can be summarized as follows. The aim is to improve the performance perceived by the user of a mobile end device. Specific targets are to reduce power consumption, increase throughput and availability, and reduce delay and jitter. The device uses a wireless MAC protocol from the IEEE 802.11 family to access a WLAN, end-to-end TCP as transport protocol for the appropriate kind of data, and an alternative protocol such as UDP for streaming data. A central point is that all performance improvements are to be attained by modifying only the mobile device, since this allows the improvements to function even when roaming into arbitrary WLANs. The wireless base station and TCP server are assumed to be compliant with the respective protocols, but otherwise to have no special features. Under these assumptions, the data transfer must be optimized separately for the uplink and downlink. To help develop suitable methods, the protocol layers on the mobile device are allowed to interact vertically where this leads to substantial improvement. Specifically, this permits optimizing the behavior for TCP data streams without worrying about possible negative consequences on other types of data.

It is well known that losses on a wireless channel have a detrimental effect on the throughput when TCP is used as transport protocol. Before considering specific up- and downlink strategies, simulations were used to quantify the effect of the network delay on this problem for the system of interest here. The results show that the severity of the problem depends strongly on this delay and that short network delays partly defuse the situation.

For the optimization of the uplink, the most logical approach is along the lines of adaptive link layer strategies, since the mobile device has control over transmission in this case. One task is then to implement these ideas efficiently for the specific MAC protocol used. In this context this article presented simulation results on the effect of MAC-layer error handling and fragmentation for the IEEE 802.11 and 11b standards. A special point of emphasis is that for TCP data, the optimal MAC behavior is a "never give up" strategy. Hereby the MAC layer never drops a TCP packet but instead tries to use a spectrum of different methods to ensure successful transmission when the channel is bad. Should transmission be impossible for a given packet, the correct response is to notify TCP that the

connection should be terminated, but not to drop the packet and continue with the next one. Also, inspection of the effect of fragmentation shows that, for these protocols, a fixed choice of packet length gives reasonably good results for a wide range of bit error rates. For the downlink, alternative approaches are still needed since the mobile device must cope with whatever the base station sends out. Based on ongoing work, the authors are convinced that suitable solutions can be found in this case also.

In conclusion, it is shown that the combination of specific separate optimization strategies for the up- and downlink with vertical interaction between the protocol layers can lead to a large improvement in the performance of a mobile terminal. As a major benefit, it is then possible to optimize the system performance without any modifications to the base stations or to the TCP protocol on the remote hosts.

References

[1] V. Tsaoussidis, H. Badr, and R. Verma, "Wait and Wave Protocol (WWP): An Energy Saving Transport Protocol for Mobile IP Devices," *Proc. ICNP*, 1999.

[2] H. Balakrishnan, S. Seshan, and R. H. Katz., "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," *WL Nets.*, vol. 1, 1995, pp. 469–81.

[3] C. Parsa and J. J. Garcia-Luna-Aceves: "Improving TCP Performance over Wireless Networks at the Link Layer," *Mobile Nets. and Apps.*, vol. 5, 2000, pp. 57–71.

[4] P. Lettieri and M. B. Srivastava, "Adaptive Frame Length Control for Wireless Link Throughput, Range, and Energy Efficiency," *Proc. IEEE INFOCOM*, New York, NY, 1998, pp. 564–71.

[5] D. Eckhardt and P. Steenkiste, "Improving Wireless LAN Performance via Adaptive Local Error Control," *Proc. ICNP*, 1998.

[6] J. Pan, J. W. Mark, S. X. Shen, "TCP Performance and Behaviors with Local Retransmissions," to appear, *J. Supercomp.*

[7] G. Xylomenos and G. C. Polyzos, "Link Layer Support for Quality of Service on Wireless Internet Links," *IEEE Pers. Commun.*, Oct. 1999.

[8] IEEE 802.11, "IEEE Standard for Wireless LAN Medium Access Control and Physical Layer Specification," 1997.

[9] K.F. Dombrowski *et al.*, "Vertical Optimization of Data Transmission for Energy Aware Mobile Devices," to appear, *Proc. IEEE Int'l. Conf. WLANs and Home Nets.*, 2001.

[10] A. DeSimonis and S. Nanda, "Wireless Data: Systems, Standards, Services," *ACM WL Nets.*, vol. 1, no. 3, 1995.

[11] H. Balakrishnan *et al.*, "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links," *IEEE/ACM Trans. Net.*, Dec. 1997.

[12] M. Zorzi and R. R. Rao, "Error Control and Energy Consumption in Communication for Nomadic Computing," *IEEE Trans. Comp.*, 1997.

[13] R. Kraemer and M. Methfessel, "A Vertical Approach to Energy Management," *Proc. MoCu*, 1999.

[14] P. Langendörfer *et al.*, "Shielding TCP from Wireless Link Errors: Retransmission Effort and Fragmentation," to appear, *J. Supercomp.*.

[15] P. Mähönen *et al.*, "Platform-Independent IP Transmission over Wireless Networks: The WINE Approach," *IEEE Pers. Commun.*, vol. 8, no. 6, 2001.

Biographies

MICHAEL METHFESSEL (methfessel@ihp-microelectronics.com) received his Ph. D. in theoretical physics in 1986 from the University of Nijmegen, The Netherlands. After some years in computational solid state physics he is now working in the systems department of the IHP in Frankfurt, Oder, with emphasis on TCP over wireless links.

KAI F. DOMBROWSKI (dombro@ihp-microelectronics.com) received his B. Sc. degree in physics at the University of Freiburg, Germany, in 1996 and his Ph. D. in physics at the Technical University of Cottbus, Germany, in 2000. His research areas since 2000 include profiling and power optimization of wireless protocols.

PETER LANGENDÖRFER (langend@ihp-microelectronics.com) received his diploma in computer science from the Technical University of Braunschweig in 1995 and his Ph. D. from the Brandenburg University of Technology at Cottbus (BTU) in 2001. From 1995 until 2000 he worked as a scientific assistant in the Department of Computer Science of the BTU. He is now with the IHP in Frankfurt, Oder. His research interests include mobile communication, protocol engineering, and automated protocol implementation.

HORST FRANKENFELDT (frankenfeldt@ihp-microelectronics.com) received his diploma in physics in 1972 from the Technical University of Madgeburg, Germany. Until 1997 he worked mainly in R&D of electron beam lithography. He is now working in the systems department of the IHP in Frankfurt, Oder, in research areas including power optimization of wireless protocols.

IRINA BABANSKAJA (babanskaja@ihp-microelectronics.com) received her diploma and Ph. D. degrees in chemistry from Moscow Lomonosov University, Russia. Until 1997 she was mainly involved in materials research and diagnostics for CMOS and SiGe technology. Her current research areas include protocol development for the wireless Internet.

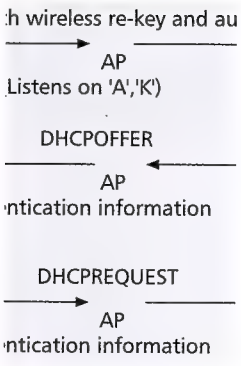
IRINA MATTHAEI (matthaei@ihp-microelectronics.com) graduated from the College of Engineering Ehemistry at Madgeburg, Germany, as an engineer of chemical technology. From 1981 to 1997 she worked in high-temperature cleanroom technology. Currently she is working in the systems department at the IHP, mainly in power optimization of wireless protocols.

ROLF KRAEMER (kraemer@ihp-microelectronics.com) received his diploma and Dr.-Ing degrees from the computer science department of the RWTH, Aachen, Germany. He has worked for 15 years in R&D of communication and multimedia systems at Philips Research in Hamburg and Aachen. Since 1998 he is professor of systems at the IHP in Frankfurt, Oder, and the BTU in Cottbus, Germany. He leads the systems research department of the IHP where his research focus is on wireless Internet systems from application to systems on chip. He is co-founder of the startup company lesswire AG where he holds the position of CTO. He is a member of the IEEE Computer Society, VDE-NTG, and German Informatics Society.

A dynamic adaptation of the transmission characteristics inherently involves complicated trade-offs. One can increase the error correction only at the price of a lower data rate, and fragment the packets only at the price of increased overhead.

YOUR 802.11 WIRELESS NETWORK HAS NO CLOTHES

WILLIAM A. ARBAUGH, NARENDAR SHANKAR, AND Y. C. JUSTIN WAN
UNIVERSITY OF MARYLAND
KAN ZHANG, HEWLETT-PACKARD LABORATORIES



Many firms, based on vendor literature, believe that the security provided by their deployed wireless access points is sufficient to prevent unauthorized access and use. Unfortunately, nothing could be further from the truth.

[1] *This article was written before the IEEE 802.11 Task Group on Security (TGi) began significant changes to WEP, and before products supporting IEEE 802.1X and EAP/TLS were available.*

ABSTRACT

The explosive growth in wireless networks over the last few years resembles the rapid growth of the Internet within the last decade. To protect internal resources, organizations usually purchased and installed an Internet firewall. We believe that the currently deployed wireless access points present a larger security problem than the early Internet connections. A large number of organizations, based on vendor literature, believe that the security provided by their deployed wireless access points is sufficient to prevent unauthorized access and use. Unfortunately, nothing could be further from the truth. While the current access points provide several security mechanisms, our work combined with the work of others show that *all* of these mechanisms are completely ineffective. As a result, organizations with deployed wireless networks are vulnerable to unauthorized use of, and access to, their internal infrastructure. In this article we present a novel solution that requires no changes or additions to any deployed wireless equipment, and is easily deployed and transparent to end users.

INTRODUCTION

Organizations are rapidly deploying wireless infrastructures based on the IEEE 802.11 standard [1]. Unfortunately, the 802.11 standard provides only limited support for confidentiality through the Wired Equivalent Privacy (WEP) protocol, which contains significant flaws in design [2–5]. Furthermore, the standards committee for 802.11 left many difficult security issues such as key management and a robust authentication mechanism as open problems. As a result, many of the organizations deploying wireless networks use either a permanent fixed cryptographic variable, or key, or no encryption whatsoever. This fact, coupled with the fact that wireless networks provide a network access point for an adversary (potentially beyond the physical security controls of the organization), creates a significant long-term security problem. Compounding this is the fact that the access control mechanisms available with currently deployed access points[1] contain serious flaws; an adversary can easily subvert them.

Organizations over the last few years have expended considerable effort to protect their internal infrastructures from external compromise. As a result, organizations have canalized their external network traffic through distinct openings protected by firewalls. The idea is simple. By limiting external connections to a few well-protected openings, the organization can better protect itself. Unfortunately, the deployment of a wireless network opens a "back door" into the internal network that permits an attacker access beyond the physical security perimeter of the organization. As a result, the attacker can implement the "parking lot" attack, where the attacker sits in the organization's parking lot and accesses hosts on the internal network. Ironically, in some cases, the existence of the firewall may make the organization's hosts more vulnerable to the attacker because of the mistaken premise that the internal hosts are immune from attack and potential compromise. Finding an effective solution to these problems is difficult since all of the security mechanisms are contained in the firmware of the wireless equipment (e.g., PCMCIA card and access points), and virtual private networks (VPNs) are not realizable in all environments.

This article describes the flaws in the two access control mechanisms (*MAC address control lists* and the proprietary *Closed Network scheme*) that exist in some access points and a simple eavesdropping attack against the 802.11 specified shared key authentication mechanism. Exploiting these flaws, when encryption is not enabled, permits an adversary immediate access to the wireless network as well as the organization's local area network. The use of encryption prevents an adversary from gaining immediate access, but combining our attacks with the weaknesses found in WEP by others provides such access [4].

The next section presents a short overview of the network configuration modes supported in the 802.11 wireless standard. This is followed by an overview of the 802.11 security mechanisms and a proprietary extension for access control. The next section describes attacks against the only two access control mechanisms available in most current access points, and an attack against the 802.11 standard shared key authentication

1536-1284/02/$17.00 © 2002 IEEE

mechanism. Finally, we conclude the article with a discussion of a higher-layer authentication and key management scheme that can easily be applied to almost any currently deployed wireless network without requiring the purchase of new hardware or the installation of new firmware in all of the organizations wireless devices.

802.11 Wireless Networks

802.11 wireless networks operate in one of two modes: *ad hoc* or *infrastructure*. The IEEE standard defines the ad hoc mode as independent basic service set (IBSS) and the infrastructure mode as basic service set (BSS). In the remainder of this section, we explain the differences between the two modes and how they operate.

In ad hoc mode, each client communicates directly with the other clients within the network. Ad hoc mode is designed so that only clients within transmission range of each other (within the same cell) can communicate. If a client in an ad hoc network wishes to communicate outside of the cell, a member of the cell must operate as a gateway and perform routing.

In infrastructure mode, each client sends all of its communication to a central station or access point (AP). The access point acts as an Ethernet bridge and forwards the communications onto the appropriate network: either wired or wireless.

Prior to communicating data, wireless clients and APs must establish a relationship, or an association. Only after an association is established can two wireless stations exchange data. In infrastructure mode, the clients associate with an AP. The association process is a two-step process involving three states:

- Unauthenticated and unassociated
- Authenticated and unassociated
- Authenticated and associated

To transition between the states, the communicating parties exchange messages called *management frames*.

A client finds a network using a very simple procedure. All APs transmit a *beacon* management frame at fixed interval. To associate with an AP and join a BSS, a client listens for beacon messages to identify the APs within range. The client then selects the BSS to join in a vendor-independent manner. For instance, on the Apple Macintosh, all of the network names (or service set identifiers, SSIDs) usually contained in the beacon frame are presented to the user so that they may select the network to join. A client may also send a probe request management frame to find an AP affiliated with a desired SSID. After identifying an AP, the client and AP perform mutual authentication by exchanging several management frames. The two standardized authentication mechanisms are described later. After successful authentication, the client moves into the second state, *authenticated and unassociated*. Moving from the second state to the third and final state, *authenticated and associated*, involves the client sending an association request frame, and the AP responding with an association response frame. After following this process, the client becomes a peer on the wireless network.

802.11 Standard Security Mechanisms

The 802.11 standard provides several mechanisms intended to provide a secure operating environment. In this section we describe each of these as well as the Closed Network access control mechanism.

Wired Equivalent Privacy Protocol

The WEP protocol was designed to provide confidentiality for network traffic using the wireless protocol. The details of the algorithm used for WEP are beyond the scope of this article. However, work by Walker *et al.* and Fluhrer *et al.* demonstrates that WEP, without dynamic key management, provides limited confidentiality and possible misuse of the network.

Open System Authentication

Open system authentication is the default authentication protocol for 802.11. As the name implies, open system authentication authenticates anyone who requests authentication. Essentially, it provides a null authentication process. Experimentation has shown that stations do perform mutual authentication using this method when joining a network, and our experiments show that the authentication management frames are sent in the clear even when WEP is enabled.

Shared Key Authentication

Shared key authentication uses a standard challenge and response along with a shared secret key to provide authentication. The station wishing to authenticate, the initiator, sends an authentication request management frame indicating that they wish to use "shared key" authentication. The recipient of the authentication request, the responder, responds by sending an authentication management frame containing 128 octets of challenge text to the initiator. The challenge text is generated by using the WEP pseudo-random number generator (PRNG) with the "shared secret" and a random initialization vector (IV). The IV is always sent in the clear as part of a WEP protected frame. Once the initiator receives the management frame from the responder, they copy the contents of the challenge text into a new management frame body. This new management frame body is then encrypted with WEP using the "shared secret" along with a new IV selected by the initiator. The encrypted management frame is then sent to the responder. The responder decrypts the received frame and verifies that the 32-bit cyclic redundancy check (CRC) integrity check value (ICV) is valid, and the challenge text matches that sent in the first message. If they do, authentication is successful. If the authentication is successful, the initiator and responder switch roles and repeat the process to ensure mutual authentication.

The value of the status code field is set to zero when successful and to an error value if unsuccessful. The element identifier identifies that the challenge text is included. The length field identifies the length of the challenge text and is fixed at 128. The challenge text includes the random challenge string.

Open system authentication is the default authentication protocol for 802.11. As the name implies, open system authentication authenticates anyone who requests authentication. Essentially, it provides a null authentication process.

Key management is a misnomer with respect to 802.11 as it is left as an exercise for vendors. As a result, only a few of the major vendors have implemented any form of key management or key agreement in their high-end products.

Closed Network Access Control

One vendor has defined a proprietary access control mechanism called "Closed Network" [5]. With this mechanism, a network manager can use either an open or a closed network. In an open network, anyone is permitted to join the network. In a Closed Network, only those clients with knowledge of the network name, or SSID, can join. In essence, the network name acts as a shared secret.

Access Control Lists

Another mechanism used by vendors (but not defined in the standard) to provide security is the use of access control lists based on the Ethernet MAC address of the client. Each AP can limit the clients of the network to those using a listed MAC address. If a client's MAC address is listed, they are permitted access to the network. If the address is not listed, access to the network is prevented.

Key Management

Key management is a misnomer with respect to 802.11 since it is left as an exercise for vendors. As a result, only a few of the major vendors have implemented any form of key management or key agreement in their high-end products. Unfortunately, none of the vendors provide sufficient information to determine the level of assurance provided by their product.

The 802.11 standard does, however, provide for two methods for using WEP keys. The first provides a window of four keys. A station or AP can decrypt packets enciphered with any one of the four keys. Transmission, however, is limited to one of the four manually entered keys — the *default* key. The second method is called a key mappings table. In this method, each unique MAC address can have a separate key. The size of a "key mappings table" should be at least ten entries according to the 802.11 specification. The maximum size, however, is likely chip-set dependent. The use of a separate key for each user mitigates the cryptographic attacks found by others, but enforcing a reasonable key period remains a problem as the keys can only be changed manually.[2]

Weaknesses in Current Access Control Mechanisms

This section describes the weaknesses in the access control mechanisms of currently deployed wireless network APs.

Closed Network Access Control Mechanism

In practice, security mechanisms based on a shared secret are robust provided the secrets are well protected in use and when distributed. Unfortunately, this is not the case with a proprietary access control mechanism. Several management messages contain the network name, or SSID, and these messages are broadcast in the clear by APs and clients. The actual management message containing the SSID depends on the vendor of the AP and the firmware version. The result, however, is that an attacker can easily sniff the network name, determining the shared secret and gaining access to the "protected" network. This flaw exists even with WEP enabled because the management messages are broadcast in the clear.

Ethernet MAC Address Access Control Lists

In theory, access control lists provide a reasonable level of security when a strong form of identity is used. Unfortunately, this is not the case with MAC addresses for two reasons. First, MAC addresses are easily sniffed by an attacker since they *must* appear in the clear even when WEP is enabled, and second, most of the wireless cards permit the changing of their MAC address via software. As a result, an attacker can easily determine the MAC addresses permitted access via eavesdropping, and then subsequently masquerade as a valid address by programming the desired address into the wireless card, bypassing the access control and gaining access to the "protected" network.

Shared Key Authentication Flaw

The current protocol for shared key authentication is easily exploited through a passive attack by the eavesdropping of one leg of a mutual authentication. The attack works because of the fixed structure of the protocol (the only difference between different authentication messages is the random challenge), and previously reported weaknesses in WEP.

The attacker first captures the second and third management messages from an authentication exchange. The second message contains the random challenge in the clear, and the third message contains the challenge encrypted with the shared authentication key. Because the attacker now knows the random challenge R, the encrypted challenge *ciphertext*, C, and the public IV, the attacker can derive the pseudo-random stream produced using WEP, $WEP_{PR}^{K,IV}$, with the shared key, K, and the public initialization variable, IV, using Eq. 1.

$$WEP_{PR}^{K,IV} = C \oplus R \quad (1)$$

The size of the recovered pseudo-random stream will be the size of the authentication frame, because all elements of the frame are known: algorithm number, sequence number, status code, element id, length, and challenge text. Furthermore, all but the challenge text will remain the same for *all* authentication responses. The attacker now has all of the elements to successfully authenticate to the target network, without knowing the shared secret K.

The attacker requests authentication of the AP it wishes to associate/join. The AP responds with an authentication challenge in the clear. The attacker then takes the random challenge text, R, and the pseudo-random stream, $WEP_{PR}^{K,IV}$, and computes a valid authentication response frame body by *XOR-ing* the two values together using Eq. 2.

$$C = WEP_{PR}^{K,IV} \oplus R \quad (2)$$

The attacker then computes a new ICV as described in Borisov *et al.* Now, the attacker responds with a valid authentication response message, and he/she associates with the AP and joins the network.

[2] *It must be noted that with the advent of 802.1X, this is no longer true. However, the TGi continues to debate the details of how key management will work within the 802.1X framework. As a result, current support for rekeying using 802.1X is not widely available.*

New Vendor Solutions and New Standards

Several proprietary solutions have been released recently. The advantage of these solutions is that the link layer key is renewed on a per-user per-session basis. However, the key does not change within the session. Even when the user session lasts for weeks, the link layer key does not change. In other words, there is no *timed key management protocol*. Also, key management is tied to authentication, because the keys are renewed only when the user authenticates. In addition, these solutions do not provide a proper means of authenticating a *disconnected user*. A disconnected user is an authorized user of the network but has left the network temporarily. As shown earlier, it is mandatory to have WEP enabled to thwart some of the attacks described by us. If WEP is enabled, such a user cannot authenticate because he/she might not have the correct window of WEP keys.

Fortunately, the 802.11 standards body is currently working on significant improvements to the standard. But all of the changes proposed by the standards body require firmware updates for all clients and APs. Furthermore, the changes are not guaranteed to work with all vendor implementations; processing power or other hardware limitations may prevent the vendor from fully implementing the proposed changes.

The task group on security (TGi) for the 802.11 standards body has been working for the past few months on an interim solution to the known problems with WEP. Their approach is to design a solution that can be implemented in the firmware of the medium access controller chip of most (not all) vendors' chipsets, or through the use of the host processor (client or AP). There are four key elements to the current draft design:

- Dynamic key management
- The addition of message integrity via a message authentication code
- Restructuring the manner in which the initialization vector and the key are combined to avoid the weak keys found by Fluhrer *et al.*
- Defining a new cipher (AES), and specifying MAC message format changes to support upper layer authentication and cipher suite negotiation

The exact details of the current TGi proposal are beyond the scope of this article, and change frequently as the draft converges to a standard [6].

Insight into our Solution

We now describe a solution we have developed at the University of Maryland. Our solution uses DHCP options as a transport mechanism for wireless key management and authentication, provides timed key management, and solves the disconnected user problem.

Design Goals

When we began this work, solutions designed to mitigate the known attacks against WiFi-based networks were proprietary or required the purchase of additional hardware. Our motivation was to find a viable solution that could be easily deployed within an enterprise without requiring the purchase and deployment of additional hardware. Our solution, proposed here, while not optimal, *mitigates* all of the known attacks. We realize that our approach *does not prevent* all of the known attacks since only low-level protocol changes by vendors and the IEEE can prevent all of the known attacks. Our solution provides a significant increase in protection for those organizations that currently cannot obtain vendor-provided upgrades (not all vendor equipment can be upgraded) and do not wish to purchase new equipment.

We have four primary design goals:

- Provide a robust key management for wireless LANs using the DHCP services of the wired LAN
- Work with the existing infrastructure
- Limit complexity (complexity breeds design and implementation errors)
- Solve the timed key management and disconnected user problems

The first goal mitigates most of the known attacks against WEP (MIC attacks cannot be prevented). The second, third and fourth goals not only provide transparency of the security mechanisms to the users, but also reduce the potential for errors in design, implementation, and operation.

Solution Overview

Retrofitting security is never a wise idea. Unfortunately, security architects often find it necessary to do so. The approach used in our solution is to limit the scope of the changes required in a wireless deployment. Therefore, we opted to make minor changes to the infrastructure rather than requiring changes to the APs and the users' wireless cards. In our solution, we:

- Use DHCP authentication for higher-layer authentication [7].
- Use a wireless rekey DHCP option for key management, which when set along with the authentication option can be used to transport the WEP key that is encrypted with a key generated from authentication.
- Solve the disconnected user problem by using a *two-door approach* at the link layer (using two keys, one a long-term WEP key used for entry into the network, and subsequently to authenticate, and another a short-term WEP key, which is used for encrypting wireless traffic). The relatively long-lived WEP key is denoted A, and the key used for communication is denoted K.
- Provide *timed key management* by tying key management and DHCP leases. When a user/station renews its IP address the corresponding WEP key is also renewed (if the wireless rekey option is set).

Key Management Protocol

The wireless network consists of stations (STAs) trying to connect to a wired network using the AP. The DHCP server exists on the wired portion of the network. All link layer traffic is encrypted using K (the current link layer key). Wireless traffic from the STAs is encrypted using K. The AP decrypts the wireless traffic

Retrofitting security is never a wise idea. Unfortunately, security architects often find it necessary to do so. The approach used in our solution is to limit the scope of the changes required in a wireless deployment.

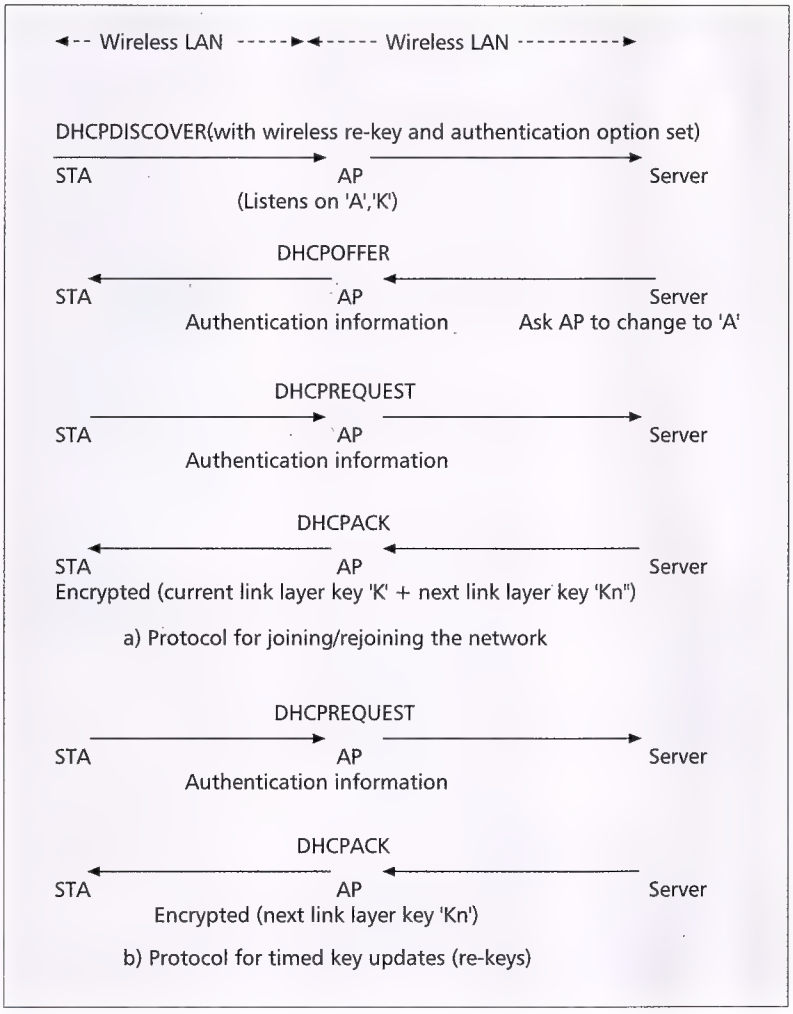


Figure 1. *DHCP authentication and rekey messages frame.*

(because it too has K) and forwards the traffic to the wired network.

The idea to mitigate the current WEP flaws is to enforce a small key period for K. At the same time valid STAs of the network, who leave the network must be able to obtain the current key (if it has changed) when they rejoin the network. This is accomplished by a *double-door entry mechanism* where the STA gains entry and authenticates itself (using higher-layer DHCP authentication) into the network using A. The frequency of use of A is significantly smaller than that of K and is hence considered to be a key with a longer lifetime. Both the AP and STAs have a window of four WEP keys. They can listen on any of the four keys but can transmit only on one key.

The AP listens on both K and A. The STAs that are a part of the network have K and A. The valid STAs who do not have K (they previously left the network and are now rejoining) can gain entry and authenticate themselves using A. After the STAs have authenticated themselves, they obtain the current link layer key K via the DHCP option.

Apart from rejoining the network, regular timed key management takes place for all STAs currently in the network by leveraging DHCP lease expirations. In other words, the current key K keeps changing frequently. We use DHCP as a *transport mechanism* for getting the new link layer key K_n.

The basic idea is as follows:

- When the client/STA joins/rejoins the network, it is assigned an IP address and given the current link layer key K. The IP address is leased for a particular time period. This time period is set by organizational policy. Apart from this, the STA is also given the next link layer WEP key, K_n.
- All the clients in the network who have the current key renew their IP address (note: the address does not need to change) depending on the lease time and in the process also obtain the new link layer key K_n.

The messages involved with the protocol are shown in Fig. 1:

- The STA sends a DHCPDISCOVER message with the wireless rekey and DHCP authentication options set. The STA transmits the message encrypted with link layer key A. The AP can listen on A and K, and hence forwards the request to the DHCP server on the wired LAN.
- The DHCP server sends a DHCPOFFER message including the authentication information in accordance with the DHCP authentication protocol [7]. We note that the AP's transmission key must be changed to A before transmission and back to K afterwards.
- The STA transmits a DHCPREQUEST message, which includes the authentication information. The authentication information is again in accordance with the DHCP authentication protocol.
- The DHCP server sends back a DHCPACK message, which includes the authentication information and encrypted current WEP key K. The encryption can be done with a shared key as defined by [7]. Also, the DHCP server sends the next WEP key K_n encrypted.

The problem of encrypting the WEP keys with the shared key as defined by [7] is that there is an inherent problem of scaling. Hence, there arises a necessity for a public-key-based authentication mechanism and the use of keys derived from the public-key-based authentication scheme.

Implementation

At the time when we were working on the implementation, none of the available versions of DHCP had the authentication option implemented. Hence, to be able to use authentication functionality in our protocol, we propose a public-key-based and a shared-key-based authentication system, which requires minimal state maintenance at the DHCP server. The public key mechanism also helps us solve the scalability problem.

The following protocols implement both authentication and rekeying functionality using the DHCP wireless rekey option.

Public Key Version — Let C denote the STA and S the DHCP server. Suppose the STA and DHCP server have public key PK_c and PK_s, respectively. The corresponding certificates are denoted by $cert_C$ and $cert_S$, respectively. For simplicity,

we assume the same public key pair is used for both confidentiality and digital signature. If different keys are desired, two pairs of public keys can be given to each STA. The following describes the protocol when an STA joins the network.

$C \Rightarrow S(DHCPDISCOVER)$

The STA sends a DHCPDISCOVER message with the wireless rekey option set. The STA transmits the message encrypted with the link layer key A. The AP can listen on both A and K, and hence forwards the request to the DHCP server on the wired LAN.

$S \Rightarrow C(DHCPOFFER)$: $nonce_S$

The DHCP server sends a DHCPOFFER message, which includes a nonce $nonce_S$ chosen by the server.

$C \Rightarrow S(DHCPREQUEST)$: $nonce_S$, $nonce_C$, $cert_C$, sig_C

Then the STA transmits a DHCPREQUEST message, which includes the server nonce, $nonce_S$, a nonce, $nonce_C$, chosen by the STA, and authentication information. The authentication information includes the X.509 certificate $cert_C$ of the STA's public key and the STA's signature sig_C on the message.

$S \Rightarrow C(DHCPACK)$: $PK_C\{K\}$,$PK_C\{K_n\}$, $nonce_C$, $nonce_{Sf}$, $cert_S$, sig_S

The DHCP server sends back a DHCPACK message which includes the current WEP key K and the next WEP key K_n encrypted with the STA's public key PK_C, the STA nonce $nonce_C$, a new server nonce $nonce_{Sf}$ for future use (when the STA renews its WEP key), and authentication information. The authentication information includes the X.509 certificate $cert_S$ of the DHCP server and the server's signature sig_S on the message.

The following protocol describes WEP key renewal for a connected STA.

$C \Rightarrow S(DHCPREQUEST)$; $nonce_S$, $nonce_C$, $cert_C$, sig_C

This message is the same as the DHCPREQUEST message sent in the above protocol, except that here the server nonce, $nonce_S$, should be the future server nonce $nonce_{Sf}$ obtained during the previous joining or renewing message exchange with the DHCP server.

$S \Rightarrow C(DHCPACK)$: $PK_C\{K_n\}$, $nonce_C$, $nonce_{Sf}$, $cert_S$, sig_S

This message is the same as the DHCPACK message sent in the above protocol, except that here the DHCP server only needs to send the next WEP key K_n encrypted with the STA's public key PK_c.

We note that the nonces used in the above protocols could very well be taken from the 64-bit replay detection field as defined by the DHCP authentication option if the option is implemented. In such a case, we do not have to store those nonce values. The specific details of the DHCP authentication option are described in [7].

Shared Key Version — Suppose the DHCP server has a master key K_m. To minimize the keys stored at the DHCP server, the shared key K_c between an STA with identification $client_{id}$ and the DHCP server is generated using a secure hash function *HMAC*.

$K_C = HMAC(K_m | client_{id} | K_m)$

The shared STA keys are distributed to STAs in an out-of-band manner.

The following describes the protocol when an STA joins the network.

$C \Rightarrow S(DHCPDISCOVER)$

The DHCPDISCOVER message stays the same as the public key version.

$S \Rightarrow C(DHCPOFFER)$: $nonce_S$

The DHCPOFFER message also stays the same. A server nonce, $nonce_S$, is sent by the server.

$C \Rightarrow S(DHCPREQUEST)$: $nonce_S$, $nonce_C$, $client_{id}$, MAC_C

The DHCPREQUEST message stays the same, except that the authentication information now consists of the STA's identifier $client_{id}$ and message authentication code MAC_c generated by the STA using the shared key K_c.

$S \Rightarrow C(DHCPACK)$: K_C, K_C, $\{K_n\}$, $nonce_C$, $nonce_{Sf}$, MAC_S

The DHCPACK message stays the same, except that:

- The WEP keys are now encrypted using shared key K_c rather than the STA's public key PK_c
- The authentication information now consists of message authentication code MACs generated by the DHCP server using the shared key K_c.

The following protocol describes WEP key renewal for a connected STA.

$C \Rightarrow S(DHCPREQUEST)$: $nonce_S$, $nonce_C$, $client_{id}$, MAC_C

The DHCPREQUEST message stays the same as the public key version, except that now the authentication information consists of the STA's identifier $client_{id}$ and message authentication code MAC_c generated by the STA using the shared key K_c. Note that, just like in the public key version of the protocol, the server nonce $nonce_S$ used in this message should be the future server nonce, $nonce_{Sf}$, obtained during the previous joining or renewing message exchange with the DHCP server.

$S \Rightarrow C(DHCPACK)$: $K_C\{K_n\}$, $nonce_C$, $nonce_{Sf}$, MAC_S

DHCPACK message stays the same, except that:

- The next WEP key K_n is now encrypted using shared key K_c rather than the STA's public key PK_c
- The authentication information now consists of message authentication code MACs generated by the DHCP server using shared key K_c.

Using a Session Key — One variant to the above protocols is to derive a session key for encrypting the WEP keys, rather than using the shared secret key K_c or the STA's public key to encrypt

To be able to use authentication functionality in our protocol, we propose a public key-based and a shared key-based authentication system, which requires minimal state maintenance at the DHCP server. The public key mechanism also helps us solve the scalability problem.

Time (ms)	Mean	Min	10% tile	Median	90% tile	Max	Std dev
Reauth of client	156	117	126	126	133	1016	119
Processing client request	28	20	26	28	29	37	2

Table 1. *The time required to obtain a new WEP key.*

the WEP keys directly. When an STA joins the network and authenticates itself using the above two join protocols, a session key K_s will be returned by the DHCP server in place of the WEP key K, together with WEP keys K and K_n encrypted using K_s.

A session key is valid for a session that, depending on key size, could be much longer than the lifetime of a WEP key. When an STA sends a DHCPREQUEST message to renew its WEP key within the same session, the DHCP server will simply send back the next WEP key encrypted with the current session key K_s. There is no authentication involved within the same session. Beyond the current session, the STA has to use the join protocol to reauthenticate itself and get a new session key. In this way, we separate authentication and session management (using different keys). It is also much cheaper for the public key version. The drawback is that now the DHCP server has to store a session key K for each STA.

A Simple Improvement — A simple improvement over the current WEP implementations would be to have a setup where:

- The DHCP server gives all valid STAs the same master secret S_m.
- The DHCP server just includes a nonce N with each lease that allows the derivation of the next WEP key $K_n = \text{hash}(S_m, N)\}$.

The above protocol uses *implicit* authentication since any valid STA would have S_m. It does, of course, suffer from all of the manual key management problems as far as S_m goes, but it is a simple protocol that requires no per STA state to be stored at the DHCP server.

Implementation Details

We now provide the details of the implementation of one of the protocols described above (the public key version).

Driver and Wireless Extension Details

We implemented a prototype on Linux using Cryptlib as our security toolkit. We used the GPL driver for the WaveLAN IEEE/ORiNOCO (maintained by Andreas Neuhaus), which was included in the pcmcia-cs-3.1.15 package for Linux. The driver fully supports the wireless extensions v9 (note that the current driver is v1.0.6, included in pcmcia-cs-3.1.24). Wireless extension support is required to change keys on the fly.

Linux wireless tools v20 provides tools, such as iwconfig and iwspy, which can configure WEP keys on the fly if the driver fully supports wireless extension. We extracted the code in iwconfig as a C function call to change the WEP key.

Client and Server Daemons

We instrumented the DHCP client and server code (v 3.0rc1pl1 from Internet Software Consortium, ISC) and changed the code to include the facility for large options. We also added an option for wireless rekeying.

Dealing with Large Options

DHCP options cannot be more than 256 bytes. Since we implemented our own authentication mechanism within the wireless rekey option, we had to deal with the large option size, which exceeds 256 bytes. In such cases, the large option was split into multiple buffers, which are logically grouped into an aggregate buffer. Large DHCP options were stored in the DHCP packet in three separate portions of the packet. These are the optional parameters field, the sname field, and the file field.

Performance Analysis

A server machine, which acted as an AP, ran our firewall and the DHCP server daemon. Another machine, which acted as a client, ran the DHCP client daemon. Both machines ran on RedHat Linux 6.2 with a Pentium III 933 MHz processor and 128 Mbytes RAM. The cards were running in the ad hoc demo mode when collecting performance data. The reason for using ad hoc mode was that it took quite some time to installing the key on the wireless card (on the AP) if the AP and the DHCP server were not on the same machine. This is because most vendors use SNMP messages for re-keying, which can take between 5 and 20 s at the worst.[3]

We implemented the public key version protocol discussed earlier, and we tested the effects of key updates upon active connections.

This was of utmost importance, because if a key update disconnected a connection, it would have been of little use. We tested our key update protocol with many types of connection-oriented protocols, such as FTP, telnet, and Netperf, to analyze the performance of the protocol. None of the active connections was adversely affected by the concurrent key update process.

We also did various types of transfers of files ranging from 200 bytes to 10 Mbytes. This range covered a number of key updates while the file was being transferred. Even when key updates were happening rapidly (once every 10 s or so) none of the connections broke.

The amount of time it took for a disconnected user to get authenticated was measured. This reflected the time needed for an STA to get the WEP key. The amount of time it took for the server to process an STA's key update request was also measured. This determined the maximum number of key renewals per second.

Table 1 shows the above two measures. These values were taken from 200 key updates. From Table 1, we see that a DHCP server can process more than 35 key update requests/s. Moreover, it only takes approximately 0.1 s for the client to reauthenticate itself and join the network.

[3] *Now it is possible to build APs running Linux, which enables us to run the DHCP server and the AP on the same machine. Even if the DHCP server and the AP run on different machines, the extra overhead for installing the key is just the network delay on the wired network, which we believe will not affect our performance numbers drastically.*

Conclusions

The combination of our results with those of Walker and Borisov *et al.* demonstrates serious flaws in all of the security mechanisms used by the vast majority of access points supporting the IEEE 802.11 wireless standard. The end result is that most of the deployed 802.11 wireless networks are at risk of compromise. An interim short-term mitigation (not a complete solution) is a robust key management system for WEP, a higher-layer authentication system, and the use of a higher-layer transport mechanism (e.g., IPSec). The combination of these mechanisms provides a robust interim solution until hardware supporting the new standards is deployed.

Acknowledgments

The authors would like to thank Jesse Walker of Intel Corporation, Mark Seiden of Securify, and Angelos Keromytis of Columbia University for providing valuable comments on a draft of this work.

References

[1] LAN MAN Standards of IEEE Comp. Soc., "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification," 1999.

[2] N. Borisov, I. Goldberg, and D. Wagner, "Intercepting Mobile Communications: The Insecurity of 802.11." *7th Annual Int'l. Conf. Mobile Comp. and Net.*, Rome, Italy, 2001.

[3] J. Walker, "Unsafe at Any Key Size: An Analysis of the WEP Encapsulation," IEEE 802.11 Task Group E, 2000.

[4] S. Fluhrer, A. Shamir, and I. Mantin, "Weaknesses in the Key Scheduling Algorithm of RC4," *Sel. Areas of Cryptography*, Toronto, Canada, 2001.

[5] W. A. Arbaugh, "An Inductive Chosen Plaintext Attack Against WEP and WEP2. 2001," IEEE 802.11 Working Group, Task Group I (Security), 2002.

[7] R. Droms, and W. Arbaugh, Authentication for DHCP Messages, 2001, Internet Engineering Task Force (IETF).

Additional Reading

[1] N. Shankar, W. Arbaugh, and K. Zhang, "A Transparent Key Management Scheme for Wireless LANs Using DHCP," HP Laboratories, Palo Alto, CA, 2001.

Biography

William Arbaugh (waa@cs.umd.edu) joined the Computer Science Department at Maryland After spending 16 years with the U.S. Department of Defense, first as a commissioned officer in the Army and then as a civilian. During those years, he served in several leadership positions in diverse areas ranging from tactical communications to advanced research in information security and networking. In his last position, he served as a senior technical advisor in an office of several hundred computer scientists, engineers, and mathematicians conducting advanced networking research and engineering. He received a B.S. from the United States Military Academy at West Point, an M.S. in computer science from Columbia University, New York City, and a Ph.D. in computer science from the University of Pennsylvania in Philadelphia. His research interests include information systems security and privacy with a focus on wireless and embedded systems and configuration management. He also currently serves on the editorial board of *IEEE Computer*, where he edits a bimonthly column on information security.

Narendar Shankar (narendar@cs.umd.edu) is a graduate student in the computer science department at the University of Maryland, College Park. His research interests are designing security mechanisms for wireless and ubiquitous computing. He is a recipient of a graduate fellowship in the computer science department.

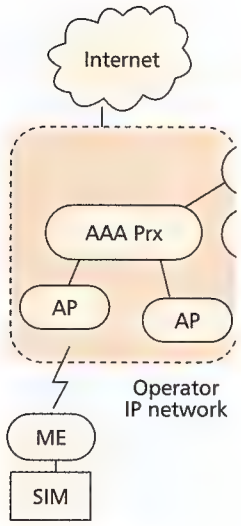
Yung-Chun Justin Wan (ycwan@cs.umd.edu) is a Ph.D. student in the Computer Science Department at the University of Maryland, College Park. He received his M.Sc. in computer science from the same school in December 2001. His current research interests are in applying algorithms to computer networks and approximation algorithms.

Kan Zhang (kzhang@hpl.hp.com) is a researcher at Hewlett-Packard Labs, Palo Alto, California. His research is focused on security and privacy issues in nomadic computing. He holds a PhD in computer science from Cambridge University.

An interim short-term mitigation (not a complete solution) is a robust key management system for WEP, a higher-layer authentication system, and the use of a higher layer transport mechanism (e.g., IPSec).

CELLULAR ACCESS CONTROL AND CHARGING FOR MOBILE OPERATOR WIRELESS LOCAL AREA NETWORKS

HENRY HAVERINEN, JOUNI MIKKONEN, AND TIMO TAKAMÄKI, NOKIA



Cellular operators, already capable of providing wide area coverage with cellular technologies, are in a good position to complement their service offerings with wireless LAN.

ABSTRACT

This article presents a system architecture, design considerations, and rationale for a mobile operator wireless LAN. The article also discusses the system implementation and performance issues. The system presented reuses GSM and GPRS mechanisms for user authentication, access control, subscriber management, operator roaming, and billing, while still being compatible with wireless Internet service provider networks and IETF and IEEE protocols such as RADIUS, EAP, and IEEE 802.1x. The architecture is a result of research carried out by Nokia between 1999 and 2002. The designed architecture has also been verified in a complete system implementation.

INTRODUCTION

Wireless local area networks (WLANs) offer high-data-rate Internet Protocol (IP) network connectivity using license-free radio frequencies. Wireless LAN coverage is becoming available in many areas, although it is generally restricted to public places and indoor premises.

This article presents an access control and charging architecture for the public access scenario, where the WLAN service provider is a cellular operator. Cellular operators, already capable of providing wide area coverage with cellular technologies, are in a good position to complement their service offerings with WLANs. Operators have a large customer base and existing systems for authentication, billing, and roaming. Moreover, the future vision of consistent user experience over various access technologies, including smooth transitions between wide area networks and WLANs, is hardly achievable unless cellular operators are involved.

This article describes a WLAN access network architecture capable of using Global System for Mobile Communications/General Packet Radio Service (GSM/GPRS) network access authorization and operator roaming in conjunction with standard IEEE 802.1x access control [1] and IEEE 802.11i WLAN security [2]. The presented architecture considers authentication, access control, and charging, with regard to wireless IP network connectivity. The architecture is a result of Nokia WLAN research conducted between 1999 and 2002. An earlier version of the architecture, described in [3, 4], has been implemented in a commercial operator WLAN solution. The solution will be updated with the architecture described in this article during 2002.

This article is structured as follows. First, the basic assumptions and requirements are discussed by comparing the solution with other proposed WLAN solutions for mobile operators. The requirements and comparison are followed by the actual architecture description. The implementation and its performance are also briefly discussed. Finally, the article concludes with considerations of the relevant standardization work and future prospects.

HIGH-LEVEL REQUIREMENTS AND COMPARISON TO OTHER SOLUTIONS

A part of the GSM success story has been written thanks to the fact that a mobile phone is always able to join the network without user intervention. The WLAN solution should also be easy to use with as little user intervention as possible. Later, when automatic transitions between WLAN and wide area networks (WANs) are supported, it will be even more important not to require user intervention upon establishment of the wireless LAN connection. One of the primary advantages cellular operators have is the ability to leverage the existing cellular user identification, roaming mechanisms, and charging solutions. Reusing the existing subscriber relationships means, in practice, that the authentication and charging should be based on the GSM Subscriber Identity Module (SIM) card. Just as a GSM phone can be put into working order by simply inserting the SIM card, the WLAN solution should not require any new credentials besides the SIM.

Not all cellular services will be available over WLAN in the first phase. The first public WLAN solutions only provide for Internet or intranet connectivity. The packet-switched cellular services, such as the services provided by the Third

1536-1284/02/$17.00 © 2002 IEEE

	User-to-network protocol	Compatible with IEEE 802.11i	Credentials required	Vendor
IEEE 802.1x with EAP/SIM (this article)	IEEE 802.1x	Yes	SIM	Nokia, Cisco
Nokia Operator Wireless LAN Release 1	NAAP	No	SIM	Nokia
SMS One Time Password with HTTP Login	HTTP over SSL	No	SIM, root public key	Ericsson and others
SMS One Time Password with IEEE 802.1x	IEEE 802.1x	Yes	SIM, root public key	Cisco?
IEEE 802.1x with GPRS Mobility Management	IEEE 802.1x	Yes	SIM, root public key	Transat

Table 1. *Operator wireless LAN solutions.*

Wireless LAN standards are being developed to better suit public access usage. Task Group I within the IEEE 802.11 Wireless LAN working group is specifying scalable network access control and enhanced security architecture, which is expected to be suitable for public Wireless LANs.

Generation Partnership Project (3GPP) IP Multimedia Subsystem (IMS) will not be available right from the beginning. Therefore, it is important that the WLAN solution allow for simultaneous cellular and WLAN sessions.

Many WLAN hotspots are operated by wireless Internet service providers (ISPs), which are not likely to deploy any "cellular overlay" protocols. Cellular operators should still be able to share WLAN radio networks with wireless ISPs. Hence, the design should be compatible with the Internet access, authorization, and accounting (AAA) networks.

WLAN standards are being developed to better suit public access usage. Task Group I within the IEEE 802.11 WLAN working group is specifying scalable network access control and enhanced security architecture, which is expected to be suitable for public WLANs. The upcoming standard will be called IEEE 802.11i, according to the task group name. The work in Task Group I is based on extensible authentication types (EAPs) and RADIUS-based authentication infrastructure. Authentication is transparent to the 802.11i access network, but the access network is able to exploit session-specific encryption and integrity protection keys. The WLAN solution for cellular operators should be compatible with IEEE 802.11i.

Table 1 includes a comparison of different WLAN solutions for mobile operators. The first row shows the architecture presented in this article. It is based on the IEEE802.1x and EAP SIM protocols. The architecture was presented publicly for the first time in an IEEE contribution [5] in January 2001, and the first version of the EAP SIM protocol, the core of the solution, was published as an Internet draft in March 2001 [6]. Other vendors later adopted the solution, and the solution is also being standardized in 3GPP for Release 6. The second row is the previous Nokia solution, described in [3, 4]. The new architecture uses a completely different set of protocols, namely IEEE 802.1x and Extensible Authentication Protocol instead of the IP-based Network Access Authentication Protocol (NAAP). The new architecture also supports identity privacy, session key distribution for packet security, and backward-compatible support for the previous solution.

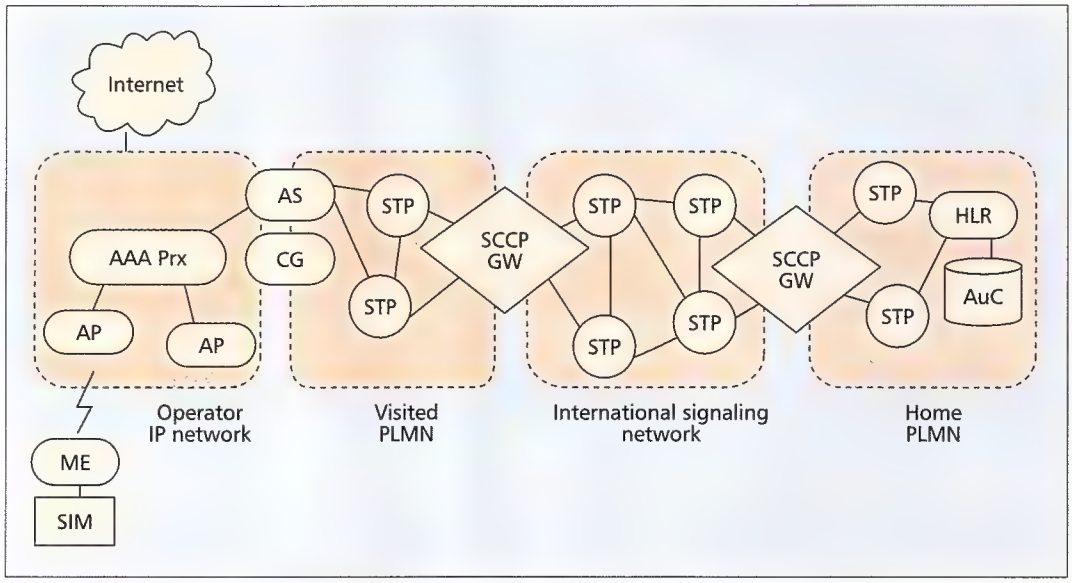
The solution on row 3, available from Ericsson makes use of one-time passwords, which are delivered with the Short Message Service (SMS) [7]. Although the solution is a little awkward to use, an advantage is that it works with a separate GSM phone and a laptop, without requiring any new software on the laptop. Because network authentication is based on the Secure Socket Layer (SSL), a suitable root public key is required on the terminal in addition to the SIM card.

The fourth and fifth solutions rely on Protected EAP (PEAP), for session key distribution and network authentication [8]. PEAP also requires root public keys for network authentication. The one-time password solution (row 4), proposed in [9], encapsulates the one-time password protocol within EAP. The last proposal (row 5) reuses GPRS mobility management messages [10]. The rationale is that the GPRS state machines and protocol implementations can be reused. However, since WLANs is currently not fully integrated with the GPRS implementations but is just an add-on network interface, these benefits cannot be achieved in current implementations. Moreover, the use of GPRS mobility management moves the user's active GPRS sessions to a WLAN, which prevents the user from having parallel cellular and WLAN sessions.

Mobile Operator Wireless LAN Architecture

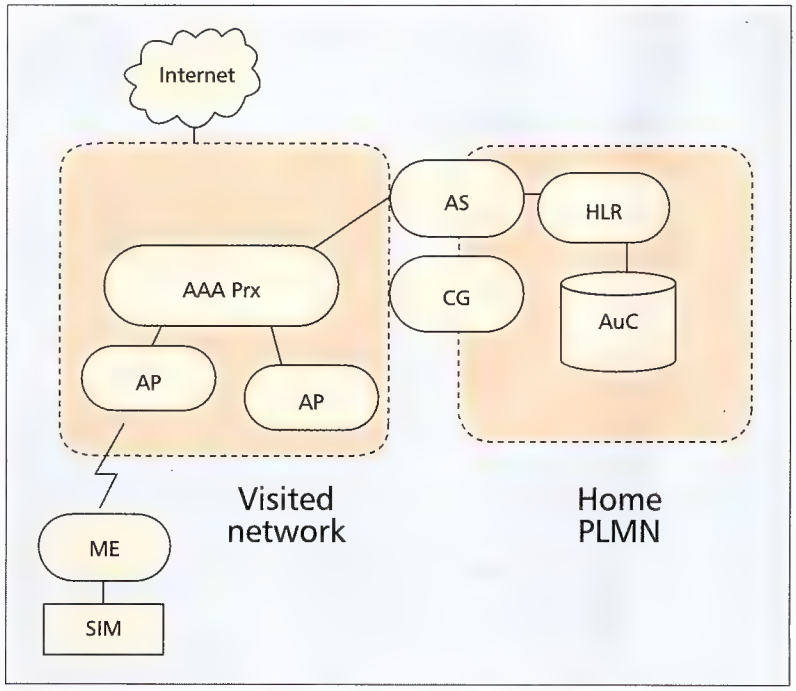
System Components

Figures 1 and 2 illustrate the architecture of the mobile operator WLAN described in this article. As it was required to support both mobile roaming infrastructure and Internet AAA infrastructure, the architecture includes an IEEE 802.11i compatible Wireless LAN access network with IEEE 802.11i access points (APs), a standard RADIUS roaming network with RADIUS proxies (AAA Prx), and the standard mobile roaming infrastructure. The GSM roaming functions are specified in various European Telecommunications Standards Institute (ETSI) and 3GPP standards, and a good and accessible description can also be found in the textbook [11]. Internet AAA operations and protocols are specified in the Internet Engineering Task Force (IETF) documents [12–17].



■ **Figure 1.** *Mobile operator WLAN architecture, GSM roaming.*

These two worlds are glued together by the authentication server (AS). The AS looks like an AAA server to the operator IP network. To the Signaling System 7 (SS7) roaming network, the AS looks like a visitor location register (VLR). As illustrated in Fig. 1, the GSM roaming network includes signaling transfer points (STP) and signaling connection control part gateways (SCCP GWs), which are able to route the AS's requests to the correct home public land mobile network (PLMN). As usual in GSM roaming, the subscriber authentication and authorization is ultimately based on the information in the home location register (HLR) and the authentication center (AuC). The WLAN system further includes a charging gateway (CG), which is responsible for forwarding the accounting records generated in the WLAN system to the operator's billing system. With AAA roaming, the interoperator protocol is RADIUS, as illustrated in Fig. 2.



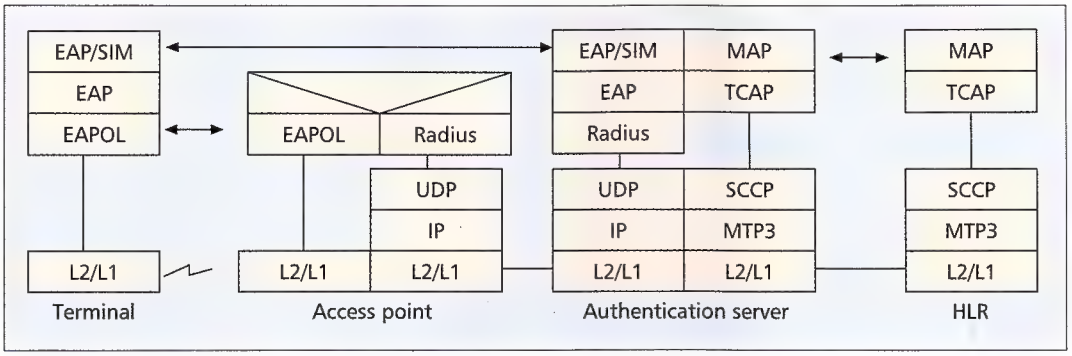
■ Figure 2. *Mobile operator WLAN architecture, AAA roaming.*

Authentication Protocol Stacks

Figure 3 shows the authentication protocol stacks used in the WLAN system. The terminal implements regular IEEE 802.1x authentication protocols. The enhanced GSM SIM authentication method is implemented as an Extensible Authentication Protocol (EAP) [16] type called EAP/SIM [18]. In IEEE 802.1x, EAP packets are encapsulated in EAPOL frames, which stands for EAP over LAN encapsulation. The notation L2/L1 used in Fig. 3 denotes the lowest two layers in the protocol stacks. In the terminal, these layers are the IEEE 802.2 and 802.11 data link layer and physical layer protocols.

Virtually any authentication method can be supported with the Extensible Authentication Protocol. EAP is a "wrapper" or an encapsulation for multi-round-trip authentication methods. The basic idea is that the terminal and the backend server perform the authentication by exchanging EAP packets, which the intermediate AAA network and AP device simply let through. The AP is only interested in the outcome of the authentication: whether the authentication succeeded or failed. The Type field in the EAP packet distinguishes different methods and hence EAP methods are often referred to as EAP types. Because the EAP authentication protocol peers are located in the terminal and in the backend server, the intermediate devices need not know the details of each authentication method.

The AP is a standard IEEE 802.11i AP and does not contain any cellular-specific functionality. The AP communicates with the backend servers using an AAA protocol. Besides authentication, accounting, and authorization, the AAA protocols also support ISP roaming operations, which allow the subscriber of an ISP to get connected using another ISP's access network. In



■ **Figure 3.** *Authentication protocol stacks.*

other words, the Internet AAA protocols are used for the same tasks as the cellular SS7 and the Mobile Application Part (MAP) protocols are used in mobile networks. At the moment, RADIUS is the most widely used AAA protocol. Diameter, the successor to RADIUS, is being standardized in the IETF.

In Fig. 3, the AP includes a RADIUS client that passes the EAP packets through to the AAA network. RADIUS runs over the User Datagram Protocol (UDP), which runs over the Internet Protocol (IP). Standard AAA proxy servers and AAA roaming, without any cellular-specific functionality, may be employed between the AP and the AS. The AS can also pass keying material derived as part of the EAP exchange to the AP to be used for packet security between the terminal and the AP.

The AS is a RADIUS server that implements the EAP/SIM authentication method peer. It also includes the Mobile Application Part protocol stack so as to be able to obtain authentication triplets and authorization information from the SS7 network. Message Transfer Part (MTP) is the SS7 network protocol that is common for all SS7 components. Message Transfer Part ensures that the signaling units are securely transferred through the network. Message Transfer Part Level 3 (MTP3) evaluates the destination point code (DPC) of each received message. If the DPC is the element's own signaling point code (SPC), MTP3 passes the message to upper layers for further processing. If the DPC of the message is not the SPC of the network element in question, MTP3 finds a suitable signaling link and sends the message further in the network to another network element. Signaling Connection Control Part (SCCP) offers connectionless and connection-oriented services for upper layer applications. Mobile Application Part uses only connectionless services for routing the messages through the signaling network. The Transaction Capabilities Application Part (TCAP) functions as an interface between SCCP and Mobile Application Part. Transaction Capabilities Application Part enables a logical exchange of messages between two network elements.

The HLR, in conjunction with the AuC, which is not shown in the figure, is the network element in the home network that sends the authentication triplets and service profile to the AS. The AuC is a database that contains a copy of the secret Ki key stored in each subscriber's SIM card.

Operation of the Operator Wireless LAN System

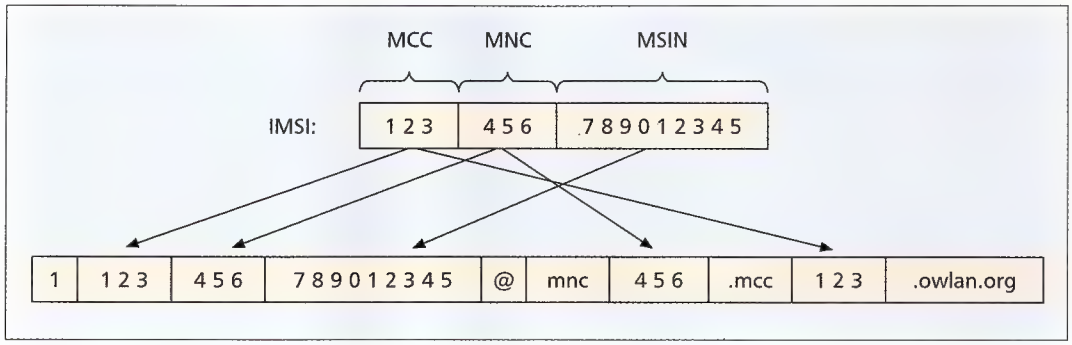
User Identity Format

In both GSM and Internet access networks, the user identity has a structured format that contains both the actual user identity and an identifier of the home network, which is used in roaming scenarios. In Internet access networks, the identity format is called a network access identifier (NAI) [17]. The NAI consists of a username portion, which is followed by an @ character and a realm portion. For example, *firstname.lastname@myisp.com* is a valid NAI. The Internet roaming network directs the AAA communications to the correct AAA server based on the realm part of the NAI. The AAA clients and proxy servers typically use their preconfigured rules to find the appropriate next hop AAA server based on the realm.

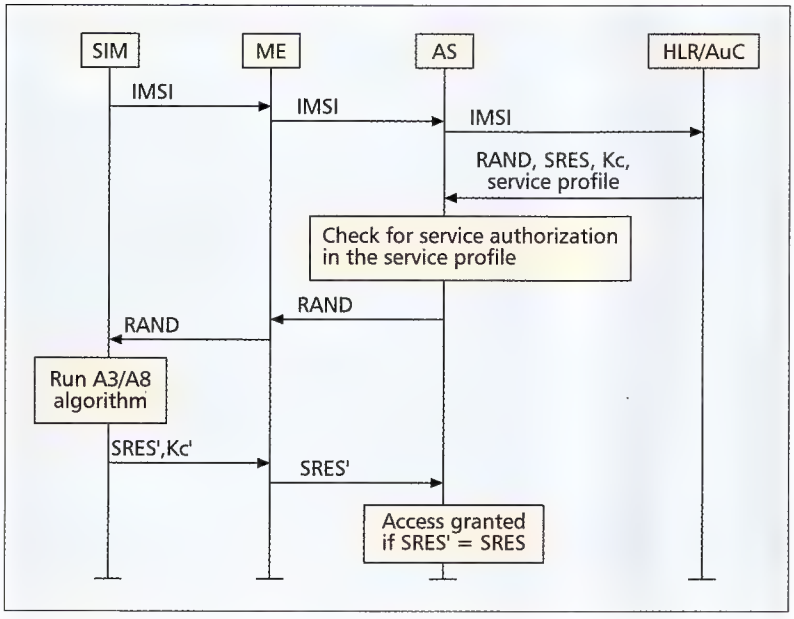
On the other hand, cellular subscribers are identified with the International Mobile Subscriber Identity (IMSI) number stored on the SIM card. The IMSI is composed of a three-digit mobile country code (MCC), a two- or three-digit mobile network code (MNC), and a not more than 10-digit Mobile Subscriber Identification Number (MSIN). The MCC and MNC uniquely identify the home GSM operator and hence can be used to find the correct HLR.

Since the present wireless LAN solution needs to be compatible with Internet access networks, the NAI format is used. The terminal maps the IMSI into an NAI by including the IMSI in the username portion of the NAI. The mapping procedure is illustrated in Fig. 4. The realm portion of the NAI can be a preconfigured string, or can be automatically derived from the MCC and MNC digits of the IMSI. For example, the IMSI *123456789012345* would be mapped to *1123456789012345@mnc456.mcc123.owlan.org*. The NAI that is derived from an IMSI always begins with character 1 to allow future developments in the NAI format. This character is followed by the IMSI. The realm part begins with *mnc* followed by the MNC, followed by *.mcc* and the MCC. Finally, the MCC is followed by *.owlan.org*, which is a realm name reserved for SIM authentication. The MNC and MCC portions in the NAI realm enable standard AAA proxy servers to forward the AP's AAA requests to the correct AuS.

The Internet AAA protocols are used for the same tasks as the cellular SS7 and the MAP protocols are used in mobile networks. At the moment, RADIUS is the most widely used AAA protocol. Diameter, the successor to RADIUS, is being standardized in the IETF.



■ **Figure 4.** *Mapping an IMSI to a network access identifier.*



■ Figure 5. *Network access authorization in GSM.*

Authentication in GSM

Figure 5 illustrates network access authorization in GSM. Authentication in GSM and GPRS is based on a challenge-response mechanism. The authentication algorithm that runs on the SIM can be given a 128-bit random number (RAND) as a challenge. The SIM runs two operator-specific confidential algorithms called A3 and A8, which take the RAND challenge and the secret Ki key as input, and produce a 32-bit response (SRES) and a 64-bit-long encryption key Kc as output, respectively. The authentication parameters RAND, SRES, and Kc are referred to as a GSM triplet.

Authentication and Roaming in the Wireless LAN System

The message sequence chart in Fig. 6 illustrates the operation of the WLAN system in a successful authentication exchange. In step 1, the AP transmits an EAP over LAN frame that contains an EAP-Request packet of the type Identity. This packet is used in EAP to request the peer to send its identity. The terminal responds with an EAP-Response/Identity that contains an NAI (step 2). As discussed above, the request is routed to the correct AS based on the realm portion of the NAI.

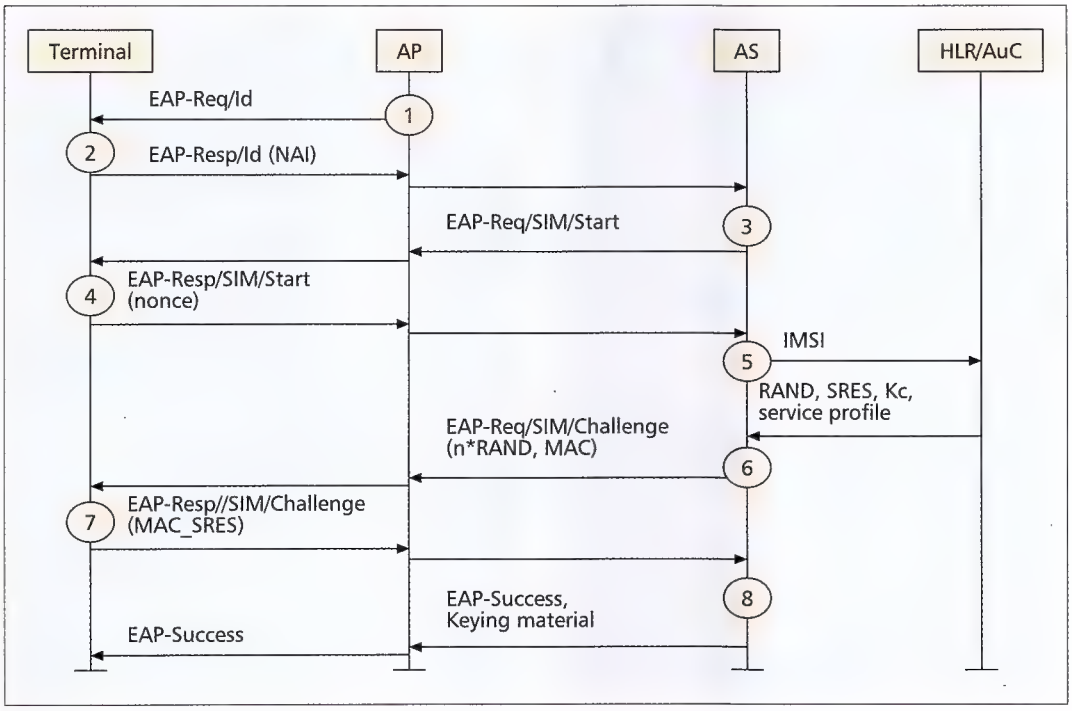
The AS starts the SIM authentication sequence by issuing the EAP-Request/ SIM/Start packet (step 3). This packet is usually a null packet but it is sometimes used in the IMSI privacy implementation described later.

In step 4, the mobile equipment chooses a long random number (nonce), which is included in the EAP-Response/SIM/Start packet. The nonce parameter can be perceived as the terminal's challenge to the network. The first EAP/SIM round-trip is required to pass the nonce parameter to the AS, since the Identity packet in EAP cannot contain any other data besides the user identity.

The AS uses the Mobile Application Part protocol over the SS7 network to obtain the user's service profile and GSM triplets from the HLR based on the IMSI (step 5). Because the AS uses the same Mobile Application Part messages as a VLR, the roaming features of the SS7 network work exactly as they work, usually in GSM. The GSM signaling system routes the Mobile Application Part messages by means of SCCP routing. The AS needs to convert the IMSI to a global title, by which the network is able to route the Mobile Application Part messages to the correct HLR. The conversion procedure is called IMSI analysis, and it is based on three CCITT numbering plans, E.164 (ISDN number), E.212 (IMSI), and E.214. The AS maps the IMSI into an E.214 number by converting the MCC portion to an E.164 country code and the MNC portion to an E.164 national destination code. The Mobile Station Identification Number portion of the IMSI is used as it is. As a result, the E.214 number is a hybrid number consisting of E.212 and E.164 numbers. The Global Title normally consists of the E.164 country code, the national destination code, and the first two digits of the MSIN.

In step 6, the AS verifies from the service profile it has received from the HLR whether the subscriber is authorized to use the WLAN service. As the GSM specifications do not (yet) know any concept of a WLAN service, the AS must use the other service bits in an operator-specific manner to deduce whether the WLAN service can be granted. Because the service bit usage convention can vary between operators, the AS needs to be configured with the operator-specific rules of each partner operator.

In step 5, the AS only uses a small subset of



■ **Figure 6.** *WLAN authentication message sequence.*

Because the wireless LAN environment is more hostile to attacks than the GSM network, the wireless LAN solution was required to have an enhanced level of security and to support network authentication and generation of stronger session keys.

the Mobile Application Part functionality required by a VLR. For example, the AS never performs a location update. This enables the same SIM card to be used for other GSM and GPRS operations simultaneously. This is in fact the reason why the AS looks like a VLR rather than a GPRS network element such as the serving GPRS support node (SGSN) — there is no such Mobile Application Part operation that a SGSN could use to obtain GSM triplets without causing a location update at the same time. A location update would prevent the user from having WLAN sessions parallel to GPRS or circuit-switched sessions, which is certainly not desired. An example of useful parallel sessions is the ability to browse the network via WLAN interface while being engaged in an active phone call.

Because the GSM authentication and key agreement algorithms are 15 years old, they do not meet all today's requirements. The GSM challenge/response mechanism does support mutual authentication, but it only authenticates the subscriber to the network. The size of the derived keying material (56 bits) is also not sufficient for the new IEEE 802.11i packet security methods. Because the wireless LAN environment is more hostile to attacks than the GSM network, the WLAN solution was required to have an enhanced level of security and to support network authentication and generation of stronger session keys.

To derive stronger keys, several, say n, GSM triplets are used in one authentication exchange. If WLAN service can be offered for the user, the AS derives keying material from the terminal's *nonce* parameter and the *Kc* keys of the n GSM triplets. The authentication server sends the EAP-Request/SIM/Challenge packet to the terminal. This request packet includes n *RAND* challenges and a keyed message authentication code (MAC) calculated over the *RAND*s with a newly derived key. The MAC parameter is used for network authentication and it can be thought of as the network's response to the terminal's challenge (*nonce*).

Upon receipt of the EAP-Request/SIM/Challenge packet in step 7, the terminal runs the A3/A8 algorithm n times, once for each received *RAND* challenge on the SIM card. The terminal derives a copy of the keying material from the *nonce* and the *Kc* keys and calculates a copy of the MAC. If the terminal's copy of the MAC does not equal the received MAC, network authentication has failed and the terminal ignores the received EAP packet. In this case the terminal does not send any values it has calculated on the SIM card to the network. If the MAC is valid, the terminal proceeds by calculating a *MAC_SRES* parameter, which is a message authentication code over the SRES response values. The terminal includes the *MAC_SRES* in the EAP-Response/SIM/Challenge packet it sends to the AS.

When the AS receives the terminal's response, it calculates a copy of the *MAC_SRES* value and compares it to the received *MAC_SRES* in step 8. The authentication is successful if these two values match. The AS sends an EAP-Success packet to the terminal. The AAA message the AS sends in step 8 includes keying material for the AP, which will be used for air interface security.

Subscriber Identity Privacy

GSM networks protect the privacy of the subscriber identity with temporary identities. The terminal does not always identify the subscriber with the globally unique IMSI, but obtains a Temporary Mobile Subscriber Identity (TMSI) from the network and uses it instead of the

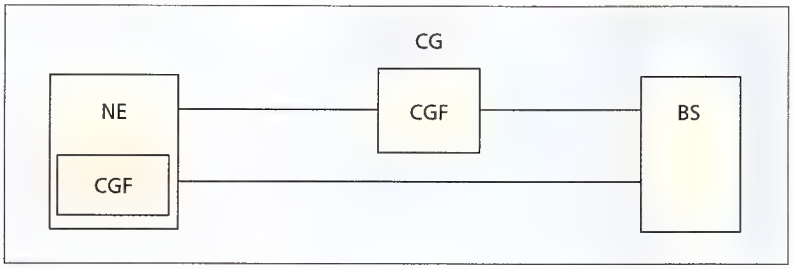


Figure 7. *The charging gateway functionality (CGF) is connected to the billing system (BS). The CGF can be implemented as a standalone charging gateway (CG) or integrated in another network element (NE).*

IMSI. The network is able to map the TMSI to the clear text IMSI. The TMSI prevents wireless observers from linking the user's activities and collecting history or profile data from the user's movements.

The WLAN architecture also protects the identity privacy with temporary identities. The TMSI cannot be directly applied, since the same subscriber may be simultaneously involved in GSM and GPRS communications that make use of the TMSI. Therefore, the WLAN system introduces a new type of temporary identities called *pseudonyms*.

In the very first connection with an AS, the client always transmits the clear text IMSI in the username portion of the NAI, as discussed earlier. In subsequent connections, the IMSI can be hidden to make the connections unlinkable to a passive eavesdropper. The EAP-Request/SIM/Challenge packet, transmitted in step 6 of Fig. 6, may include an encrypted and integrity protected pseudonym. The keying material required for the cryptographic operations are derived from the *n Kc* keys and the *nonce* value, so the correct terminal is only able to decrypt the information and obtain the new pseudonym identity. The pseudonym is a string that the terminal can use as the username portion of the NAI in the next connection. The AS only needs to be able to map the pseudonym to the clear text identity. The AS can maintain a mapping table, or it can produce pseudonyms so that the pseudonym contains the IMSI in an encrypted form. In roaming scenarios, the AS may not be able to map the pseudonym generated by another operator's AS to the IMSI. Hence, the EAP/SIM protocol includes a fallback mechanism by which the AS can request the terminal to send its clear text identity in case the server is unable to recover the IMSI from the pseudonym.

Accounting and Billing

In GPRS networks, charging is usually based either on the amount of transferred data or a fixed monthly fee. Charging data is collected and transmitted into the billing system in the form of charging data records (CDRs). The GPRS system has a concept of charging gateway functionality (CGF), which offers a means for transferring charging information reliably to the operator's billing system (BS). CGF may be integrated within a network element or a separate network device, a CG (CG), that solely concentrates on charging functions. This is illustrated in Fig. 7.

GPRS Tunneling Protocol for Charging (GTP') is an optional charging protocol that has been standardized for charging in GPRS networks. GTP' offers a reliable and redundant way of transferring CDRs from network elements to a CG. The interface from the CG to the operator's billing system is often proprietary and requires tailoring in the CG. The billing system generates the final bill to the subscriber based on the billing parameters and rules the operator has set.

GPRS charging in roaming scenarios has not yet been fully standardized. In practical solutions operators have agreed case by case with their partners on how to handle charging and utilize their existing systems.

Internet AAA protocols also include accounting functionality. The WLAN AP can collect data on the usage of resources such as connection times and amount of transferred data, and send the accounting data to the AAA server through the AAA network. Both RADIUS and Diameter include accounting extensions for transmitting accounting data. Since our WLAN architecture is compatible with standard Internet access networks, the solution supports RADIUS accounting. The accounting data the authentication server receives from the access network includes the amounts of uploaded and downloaded data traffic and session duration, as well as relevant location and identification information. The AP can send interim accounting data to the AS at certain intervals to ensure that accounting data is not lost in case of malfunction in the AP. After the terminal has left the network, the AP sends the final accounting packet to the AS indicating that the terminal is no longer using the network resources. At this point, the AS also removes the terminal entry from its internal data structures.

The AS forms the actual CDRs that are transmitted to the CG using GTP'. The CDR format used in the WLAN system closely resembles the G-CDRs used in a GPRS system. Only some fields have been modified to suit the context of WLAN charging. The AS can be configured to also send the interim charging information to the CG, thus decreasing the probability of charging data loss. Also, the interim charging information can be used later when prepaid charging features are added to the system. Once the terminal has left the network, the CG combines the possible interim CDRs and sends the final CDR to the billing system for further processing.

An alternative way of transferring the CDR to the billing system is to use File Transfer Protocol (FTP). In this case the billing system fetches final CDRs directly from the AS's hard drive, so the CG does not need to act as a mediator. However, this option is not often preferred because usually a lot of integration effort is required in the billing system in order to use the CDRs directly.

System Implementation

The architecture has been implemented in the Nokia Operator Wireless LAN solution release 2.0. The Nokia A036 AP runs the Linux operating system. It has two internal antennas that provide space and polarization diversity. The AP

includes the Intersil Prism WLAN chipset and implements the IEEE802.3af Power over Ethernet standard.

The authentication server is essentially a gateway device between the IP access network and the cellular signaling network during the terminal authentication phase. Furthermore, the authentication server also has the responsibility of delivering the charging data to the operator's CG. The AS is based on the Compaq Proliant rack mount PC hardware. It includes an Ethernet network interface and the TX3220 Mobile Application Part/SS7 interface by Natural Microsystems. The server runs the Windows NT 4.0 operating system and RADIUS server and SIM authentication software.

The current version of the AS is capable of handling approximately 10–20 terminal authentication exchanges/s depending on the configuration. Each exchange takes three roundtrips of EAP authentication messages and two SS7 Mobile Application Part messages to complete. The resulting authentication IP signaling traffic is thus 30–60 messages/s and SS7 signaling traffic 20–40 messages/s. For example, if each terminal has an average connection time of 30 min (1800 s), by using Little's formula ($\overline{N} = \lambda \overline{W}$) an AS is capable of handling 18,000–36,000 simultaneous connections in the IP access network.

The performance is also affected by the charging traffic. The AS receives the charging data from the WLAN access zone and sends it further to the operator's billing system. The AP can be configured to send interim accounting packets at certain intervals. If this interval is short, a large number of charging messages will be sent toward the core network and the AS. Therefore, the charging parameters in an access network need to be determined carefully. Considering the example above, if an AP transmits accounting information every 10 min, a network of 18,000–36,000 simultaneous users generates additional accounting traffic of 30–60 messages/s. Although the AS is capable of handling the resulting signaling traffic easily, the problem arises when the charging data is written to a hard drive. The AS can be configured to store the intermediate charging information in either its memory or hard drive. Once the data traffic due to interim charging data packets climbs above 30 messages/s the performance of the AS begins to decline as the disk write operations start to take too much time. If the data is not written to a hard drive, the interim charging data traffic is usually insignificant to the AS's performance.

During the authentication phase, the AS sends the AP a session timeout parameter that specifies how long the session is valid with the given credentials. After the session timeout period has elapsed the terminal needs to authenticate again. If this parameter value is shorter than the average session length, the performance of the system decreases accordingly. In other words, the amount of simultaneous users in the system corresponds to the value of the session timeout parameter.

For redundancy and load-balancing reasons it is recommended to have always at least two ASs in a WLAN system.

Standardization

Integration of WLAN access networks into the mobile operators' service offering is currently being studied in many standardization organizations. The 3GPP has started standardizing WLAN interworking for the Universal Mobile Telecommunications System (UMTS). 3GPP has identified several levels of interworking, ranging from common customer care and billing to offering all cellular packet-switched and even circuit-switched services over WLAN. One of the authors has been actively contributing to the WLAN interworking standardization in 3GPP, and the current draft Technical Specification resembles the present architecture very closely.

As discussed above, the present solution supports both Mobile Application Part and RADIUS roaming. However, Mobile Application Part roaming for WLAN will not be included in the 3GPP specifications, but the interoperator interface will be based on the Diameter protocol. In the absence of a global Diameter roaming network, the present architecture with its ability to reuse the SS7 network for roaming provides a viable interim solution, before RADIUS or Diameter-based roaming networks are available.

Conclusions and Future Prospects

Cellular operators will be in a position to manage services and roaming between various access networks. Service management, mobile service user trust, and global roaming are the key factors justifying the architecture presented in this article. The architecture, which supports cellular access control and subscriber management, has been implemented and is expected to be commercially available by the time this issue is published.

The solution is generic enough to be used on any access networks that support EAP. For example, EAP is one of the alternatives in the Bluetooth network access standardization. As soon as EAP and RADIUS-based authentication is supported in Bluetooth, the solution can be used to realize cellular access control and roaming for Bluetooth hotspots.

UMTS authentication and key agreement algorithms offer significant improvements over the GSM algorithms. UMTS authentication includes network authentication and strong key generation. UMTS roaming and authentication will not have a big impact on the present WLAN architecture. In fact it is sufficient to replace the EAP SIM protocol with an EAP method for UMTS authentication. One of the authors is a coauthor of [19], which is an IETF standardization contribution for EAP UMTS authentication.

Acknowledgment

Besides the authors, many people have contributed to the design of the described system and provided useful comments on the standardization contributions. Mr. Jukka Tuomi is one of the main architects of the system, and his insight has been most helpful in writing this article. In addition, the authors would like to acknowledge the substantial contributions of Juha Ala-Laurila, N. Asokan, Jukka-Pekka Honkanen, and Jyri

The authentication server is essentially a gateway device between the IP access network and the cellular signaling network during the terminal authentication phase. Furthermore, the authentication server also has the responsibility of delivering the charging data to the operator's charging gateway.

Service management, mobile service user trust, and global roaming are the key factors justifying the architecture presented in this article. The architecture, which supports cellular access control and subscriber management, has been implemented and is expected to be commercially available by the time this issue is published.

Rinnemaa. Thanks to Professor Jarmo Harju and Dr. Jukka Koskinen for their most helpful comments on a draft of this article.

References

[1] IEEE, "IEEE Standard for Local and Metropolitan Area Networks — Port-Based Network Access Control," IEEE Std 802.1X-2001, June 2001
[2] IEEE, "Draft Supplement to Standard for Telecommunications and Information Exchange Between Systems — LAN/MAN Specific Requirements — Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Specification for Enhanced Security," IEEE Std 802.11i/D2.0, Draft supplement to IEEE Std 802.11, 1999 Edition, work in progress, March 2002.
[3] J. Ala-Laurila, J. Mikkonen, and J. Rinnemaa, "Wireless LAN Access Architecture for Mobile Operators," *IEEE Communications*, vol. 39 no.11, Nov. 2001, pp. 82–89.
[4] B. Bing, *Wireless Local Area Networks — The New Wireless Revolution*, Ch. 10, Wiley., 2002.
[5] H. Haverinen and J. P. Edney, "Use of GSM SIM Authentication in IEEE802.11 System," IEEE cont., 01/039 Task Group E, Jan. 2001.
[6] H. Haverinen, "EAP SIM Authentication (Version 1)," Internet draft draft-haverinen-pppext- eap-sim-00.txt, IETF, Mar. 2001, work in progress.
[7] T. Bostrom, *et al.*, "Ericsson Mobile Operator WLAN Solution," *Ericsson Rev.*, no.1/2002, pp. 36–43.
[8] H. Andersson *et al.*, "Protected EAP Protocol (PEAP)," Internet draft draft-josefsson-pppext- eap-tls-eap-02.txt, IETF, Feb. 2002, work in progress
[9] Cisco Systems, "802.11i Secure SMS Based One Time Password Authentication," 3GPP TSG-SA WG2 mtg. #26, 3GPP doc. no. S2-022149, Aug. 2002, Toronto, Canada.
[10] A. Buckley et al., "EAP SIM GMM Authentication," Internet draft draft-buckley-pppext-eap- sim-gmm-00.txt, IETF, Aug. 2002, work in progress
[11] M. Mouly and M.-B. Pautet, "The GSM System for Mobile Communications," published by the authors, 1992.
[12] C. Rigney *et al.*, "Remote Authentication Dial In User Service (RADIUS)," IETF RFC 2865, June 2000.
[13] C. Rigney, "RADIUS Accounting," IETF RFC 2866, June 2000.
[14] C. Rigney, W. Willats, and P. Calhoun, "RADIUS Extensions," IETF RFC 2869, June 2000.
[15] B. Aboba *et al.*, "Review of Roaming Implementations," IETF RFC 2194, Sept. 1997.
[16] L. Blunk and J. Vollbrecht, "PPP Extensible Authentication Protocol (EAP)," IETF RFC 2284, Mar. 1998.
[17] B. Aboba and M. Beadles, "The Network Access Identifier," IETF RFC 2486, Jan. 1999.
[18] H. Haverinen, "EAP SIM Authentication," Internet draft draft-haverinen-pppext-eap-sim-03. txt, IETF, Feb. 2002, work in progress.
[19] J. Arkko and H. Haverinen, "EAP AKA Authentication," Internet draft draft-arkko-pppext-eap-aka- 03.txt, IETF, Feb. 2002, work in progress.

Biographies

HENRY HAVERINEN is a senior specialist at Nokia Mobile Phones, investigating mobile wireless networking, network security protocols, and the interworking of wireless LAN and cellular technologies. He currently serves as the Nokia Mobile Phones representative in 3GPP wireless LAN interworking standardization and is co-author of several related IETF documents. He has published a number of papers in the areas of wireless LANs, mobile networking, and network security. He is a Ph.D. student at Tampere University of Technology, where he received an M.Sc. in 1998.

JOUNI MIKKONEN (jouni.mikkonen@nokia.com) received his Ph.D. degree in telecommunications from the Technical University of Tampere in 1999. He has been with Nokia since 1992. He has been involved in the development of GSM network and new data services at both Nokia Telecommunication and Nokia Mobile Phones. Between September 1995 and October 1998 he worked as technical manager of the Magic WAND (Wireless ATM Network Demonstrator) project, which was the European Union funded activity within the Fourth Framework program called ACTS (Advanced Communications Technologies and Services). In April 2000, he was nominated senior technology manager of wireless broadband technologies within Nokia.

TIMO TAKAMÄKI is a competence area manager at Nokia Networks, heading a group of professionals implementing part of the Nokia Operator Wireless LAN solution, and he has actively contributed to the development of a wireless LAN and cellular interworking system. He received his M.Sc. in 2000 at Tampere University of Technology in telecommunications engineering.

ComSoc's Next Generation Digital Library

www.comsoc.org

COMMONTOLOGY

Communications Technology Examined....

An expert overview of the fields comprising communications technology can be seen by visiting the Digital Library on the IEEE Communications Society web site - www.comsoc.org.

Publishing magazines and journals and sponsoring technical conferences for professional engineers for over 50 years, the society assembled and organized it's considerable mass of publications content in a unified digital library that offers both the expert and beginning researcher a lot more information than a list of papers. The IEEE Communications Society Digital Library is organized into a discipline-specific ontology. All contributions are categorized into technical subject areas that are delineated into subspecialties.

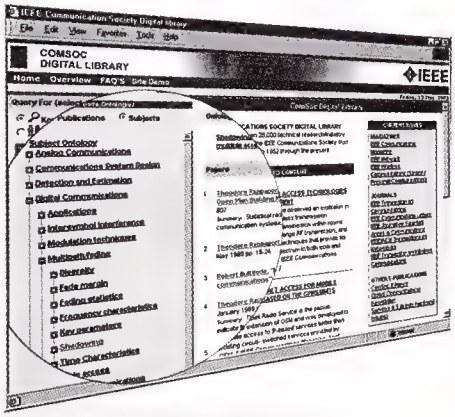
What is an Ontology?

ONTOLOGIES are hierarchical (tree-like) classifications of all the articles and people in a collection. These enable easy access to the content for novices and experts. The Publications ontology organizes all content according to the publication, volume and issue in which it appeared. The Subjects ontology organizes the collection by general topics and more specific subtopics. There are approximately 500 topics in the subject ontology, with subtopics going 4-7 levels deep. The subject ontology in particular is useful for giving an overview of the field.

The subject ontology has been created manually by experts in the field of communications, and is maintained and updated by the Communications Society's ontology review committee. Individual articles are classified into the ontology topics automatically, based on an analysis of keyword usage, references and citations to other articles in the collection, and similarity to manually identified "key papers" within each topic.

IEEE Communications Society President Celia Desmond on the CommOntology:

"This approach to our technical domain has enabled a much improved process for keyword and other meta-tag assignments to past, current, and future technical articles in the field....Our core business is to attract, aggregate, and disseminate the best communications technology intellectual property in the world."



How does it work?

Locate the paper you want in the online **ComSoc Digital Library** by a traditional word or Boolean search, ontology subject selection, or browsing author names or publication titles. An ontology selection brings you a linked list of recommended papers, a linked selection of experts, and a linked list of related topics within the ontology configuration.

Each paper's characteristics are summarized in a summary page or "mini" paper screen which contains:

- **Title:** (the name of the paper)
- **Full Text:** (link to pdf file)
- **Author:** (authors names and links to bios)
- **Bibliographic Entry:** (the publication/location and page numbers of where the paper appeared in print)
- **Ontology Subjects:** (linked ontology classifications)
- **Keywords**
- **Abstract**
- **References:** (linked references)
- **Citations:** (linked list of papers that have referenced this paper after its publication)

Take a demonstration drive...

Come to the **ComSoc Digital Library** at **www.comsoc.org** and examine the CommOntology. We think you will like what you see. Members can browse, search or access the ontology and its links. Nonmembers can request a password by sending an e-mail message to **DigitalLibrary@comsoc.org**

IEEE

IEEE COMMUNICATIONS SOCIETY

SMART HOMES

SAJAL K. DAS DIANE J. COOK

Smart homes link computers to everyday tasks and environments that have traditionally been viewed as outside the purview of automation. Important features of such environments are that they possess a degree of autonomy, adapt themselves to changing conditions, and communicate with humans in a natural way. Instead of finding application in spreadsheets and word processors, these systems take the form of intelligent rooms and personal assistants.

There are several characteristics that are commonly found in smart homes. A smart home assumes control of devices and coordinates multiple devices within the home/office, relieving the inhabitants of this burden. Interaction with smart homes is in a form that is comfortable to people. User interface primitives are not windows, icons, menus, and pointers, but include speech, gesture, and context. A smart home naturally requires ubiquitous and location-/context-aware computing, allowing the environment to process information as if computational devices are everywhere, rather than actually embedding devices everywhere. A smart home must also adapt to inhabitants and to changing conditions.

Designing and implementing smart homes often requires a unique breadth of knowledge not limited to a single discipline, but integrating aspects of machine learning, human-machine interfaces, wireless networking, and mobile communications. The purpose of this special issue is to bring together researchers working on diverse aspects of this emerging area in order to define the state of the art in smart homes, to identify fundamental issues and challenges, and to discuss ideas for further investigation.

In response to the call for papers, we received an overwhelming response. Through a rigorous review process, we have selected the following six papers for publication in this feature topic on smart homes. These articles are representative of various important aspects of smart homes, such as hardware and software requirements, computational architecture design, programming environments, user interfaces, prediction algorithms, ubiquitous and location-/context-aware computing, home networking, communication protocols, as well as intelligent device control and interdevice communication and coordination. We briefly summarize below each of these six articles.

The article by Borchers, Ringel, Tyler, and Fox, "Stanford Interactive Workspaces: A Framework for Physical and Graphical User Interface Prototyping," proposes Interactive Stuff (iStuff) as a hardware abstraction technique that enables quick customization and reconfiguration of smart home solutions. This article also proposes the Interactive Room Operating System (IROS), which creates a flexible and robust software framework that allows custom and legacy applications to communicate with each other and with user interface devices in a dynamically reconfigurable way.

In the article "Facilitating the Programming of the Smart Home," Jahnke, d'Entremont, and Stier deal with software programs that facilitate interoperability and flexibility of embedded devices in smart homes, and discuss how recent technological progress in the areas of visual programming languages and tools, component software, and connection-based programming methodologies can be applied to programming synergistic devices for smart homes.

The article by Das, Cook, Bhattacharya, Heierman, and Lin, "The Role of Prediction Algorithms in the MavHome Smart Home Architecture," presents the architecture of the Managing an Adaptive Versatile Home (MavHome) project, which is to create a smart home that acts as a rational agent. In order to maximize inhabitant comfort and minimize operation cost, the agent must be able to predict the movement and activity patterns of the inhabitants as well as their device usages. The authors propose three prediction algorithms that are central to home operations.

In their article "A Hybrid Analysis and Architectural Design Method for Development of Smart Home Components," Durrett, Burnell, and Priest identify the need for a requirement analysis of the rapidly emerging smart home technologies, and how to optimize user satisfaction considering the wide range of user types and preferences in such dynamic design environment. Based on the synthesis of several known techniques, they propose the fusion of user case and House of Quality analysis models, simulators, and prototypes.

Lin and Tseng, in "Adaptive Sniff Scheduling Scheme for Power Saving in Bluetooth," study the problem of managing the low-power sniff mode in Bluetooth, which is expected to be an important basic component of smart homes, among other portable and battery-operated devices. Based on the underlying master-slave concept of Bluetooth operations, the authors propose an adaptive protocol to dynamically adjust each slave's sniffing period in a piconet such that a trade-off is attained between the traffic and power-saving requirements.

Last but not the least, "A Power Line Communication Network Infrastructure for the Smart Home" by Lin, Latchman, and Lee suggests the use of power line LANs as a basic infrastructure for building integrated smart homes in which simple control devices to multimedia entertainment systems are seamlessly interconnected by the wires providing electricity. By simulation and actual measurements using commercial powerline products, they show that the HomePlug media access control and physical layer protocols can support delay-sensitive data streams for smart home applications.

We take this opportunity to thank all the authors of submitted papers and also the reviewers for making this feature topic truly special. We are grateful to the Editor-in-Chief, Mahmoud Naghshineh, for his encouragement and support.

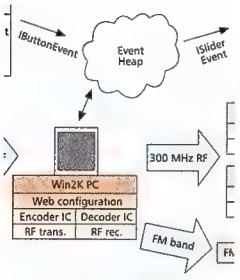
BIOGRAPHIES

SAJAL K. DAS (das@cse.uta.edu) received his B.Tech. in 1983 from Calcutta University, M.S. in 1984 from the Indian Institute of Science, and Ph.D. in 1988 from the University of Central Florida. Currently he is a Professor of Computer Science and Engineering and Director of the CReWMaN Center at the University of Texas at Arlington. His research interests include wireless networks, mobile and pervasive computing, parallel/distributed processing, performance modeling and simulation. He has published over 200 papers in these areas.

DIANE COOK (cook@cse.uta.ed) is currently a professor in the Computer Science and Engineering Department at the University of Texas at Arlington. Her research interests include artificial intelligence, machine learning, data mining, robotics, and parallel algorithms for artificial intelligence. She has published over 120 papers in these areas. She received her B.S. from Wheaton College in 1985, and her M.S. and Ph.D. from the University of Illinois in 1987 and 1990, respectively.

STANFORD INTERACTIVE WORKSPACES: A FRAMEWORK FOR PHYSICAL AND GRAPHICAL USER INTERFACE PROTOTYPING

JAN BORCHERS, MEREDITH RINGEL, JOSHUA TYLER, AND ARMANDO FOX
STANFORD UNIVERSITY



Most smart homes are created evolutionarily. This incremental addition of technology requires a highly flexible infrastructure to accommodate both future extensions and legacy systems without requiring extensive rewiring of hardware or reconfiguration on the software level.

OVERVIEW

Most smart homes are created evolutionarily by adding more and more technologies to an existing home, rather than being developed on a single occasion by building a new home from scratch. This incremental addition of technology requires a highly flexible infrastructure to accommodate both future extensions and legacy systems without requiring extensive rewiring of hardware or extensive reconfiguration on the software level. Stanford's iStuff (Interactive Stuff) provides an example of a hardware interface abstraction technique that enables quick customization and reconfiguration of Smart Home solutions. iStuff gains its power from its combination with the Stanford Interactive Room Operating System (iROS), which creates a flexible and robust software framework that allows custom and legacy applications to communicate with each other and with user interface devices in a dynamically configurable way.

The Stanford Interactive Room (iRoom,Fig. 1), while not a residential environment, has many characteristics of a smart home: a wide array of advanced user interface technologies, abundant computation power, and infrastructure with which to coordinate the use of these resources (for more information on the iRoom or the Interactive Workspaces project, please visit http://iwork.stanford.edu).As a result, many aspects of the iRoom environment have strong implications for, and can be intuitively translated to, smart homes. In particular, the rapid and fluid development of physical user interfaces using iStuff and the iROS, which has been demonstrated in the iRoom, is an equally powerful concept for designing and living in smart homes.

Before focusing on the details of iStuff, we describe the software infrastructure on which it is based and the considerations that went into designing this infrastructure.

iROS: APPLICATION COORDINATION IN UBIQUITOUS COMPUTING ENVIRONMENTS

SOFTWARE REQUIREMENTS FOR RAPID INTEGRATION AND EVOLUTION

The ability to continually integrate new technologies and handle failures in a noncatastrophic manner is essential to smart homes and related ubiquitous computing environments. Our experience working in the Stanford iRoom enables us to identify four important requirements for a software infrastructure in a ubiquitous computing environment.

Heterogeneity: The software infrastructure must accommodate a tremendous variety of devices with widely ranging capabilities. This implies that it should be lightweight and make few assumptions about client devices so that the effort to "port" any necessary software components to new devices will be small.

Robustness: The software system as a whole must be robust against transient or partial failures of particular components. Failures should not cascade, and failure or unexpected behavior of one component should not be able to infect the rest of the working system.

Evolvability: The application program interface (API) provided must be sufficiently flexible to maintain forward and backward compatibility as technology evolves. For example, it should be possible to integrate a new type of pointing device that provides higher resolution or additional features not found in older devices, without breaking compatibility with those older devices or existing applications.

Compatibility: It should be easy to leverage legacy applications and technologies as building blocks. For example, Web technologies have been used for user interface (UI) prototyping, accessing remote applications, and bringing rich content to small devices; desktop productivity

1536-1284/02/$17.00 © 2002 IEEE

applications such as Microsoft PowerPoint™ contain many elements of a "rich content display server;" and so on. Furthermore, since technology in smart spaces tends to accrete over time, today's new hardware and software will rapidly become tomorrow's legacy hardware and software, so this problem will not go away.

Our prototype meta-operating system, iROS (Interactive Room Operating System), meets the above criteria. We call it a meta-OS since it consists entirely of user-level code running on unmodified commodity operating systems, connecting the various iRoom entities into a "system of systems." We discuss the main principles of iROS here to give the reader an understanding of how it facilitates building new behaviors using iStuff.

iROS and Application Coordination

We will frame our discussion in the context of the Stanford iRoom, a prototype environment we constructed that we believe is representative of an important class of ubiquitous computing installations. The iRoom is intended to be a dedicated technology-enhanced space where people come together for collaborative problem solving (meetings, design reviews, brainstorming, etc.), and applications we prototyped and deployed were driven by such scenarios.

The basis of iROS is *application coordination*. In the original formulation of Gelernter and Carriero [1], coordination languages express the interaction between autonomous processes, and computation languages express how calculations of those processes proceed. For example, procedure calls are a special case in which the caller process suspends itself pending a response from the callee. Gelertner and Carriero argue that computation and coordination are orthogonal and that there are benefits to expressing coordination in a separate general-purpose coordination language; our problem constraint of integrating existing diverse components across heterogeneous platforms leads directly to separating computation (the existing applications themselves) from coordination (how their behaviors can be linked).

In iROS, the coordination layer is called the *Event Heap* [2].The name was chosen to reflect that its functionality could be viewed as analogous to the traditional event queue in single-computer operating systems. The Event Heap is an enhanced version of a tuplespace, one of the general-purpose coordination languages identified by Gelernter and Carriero. A tuple is a collection of ordered fields; a tuplespace is a "blackboard" visible to all participants in a particular scope (in our case, all software entities in the iRoom), in which any entity can post a tuple and any entity can retrieve or subscribe for notification of new tuples matching a wildcard-based matching template. We have identified important advantages of this coordination approach over using rendezvous and RMI (as Jini does) or simple client-server techniques (as has been done using HTTP, Tcl/Tk [3], and other approaches); these advantages include improved robustness due to decoupling of communicating entities, rapid integration of new platforms due to the extremely lightweight client API (we support all major programming languages, including HTML, for posting and retrieving tuples), and the ability to accommodate legacy applications (simple "hooks" written in Visual Basic or Java can be used to connect existing productivity, Web, and desktop applications to the iRoom).The only criterion for making a new device or application "iRoom-aware" is its ability to post and/or subscribe to tuples in the Event Heap; since we can create Web pages that do this, any device that enters the room running a Web browser is already minimally iRoom-aware.

Figure 1. *The Stanford iRoom contains a wireless GyroMouse and keyboard (visible on the table), three touch-sensitive SmartBOARDs and one non-touch-sensitive tabletop display, and a custom-built OpenGL hi-res graphic mural. The room is networked using IEEE 802.11b wireless Ethernet. Except for the hi-res mural and the tabletop, all hardware is off-the-shelf, all operating systems are unmodified Windows (various flavors) or Linux, and all software we have written is user-level.*

On-the-Fly User Interface Generation in iROS

The Event Heap is the core of iROS, but we have also built other iROS services that provide higher-level functionality. Most notably, the Interface Crafter (iCrafter) framework [4] can generate UIs dynamically for virtually any iRoom entity and on virtually any iRoom-aware device. Although it extends previous work in several important ways, including integration of service discovery with robustness and the ability to create UIs ranging from fully custom to fully automatic, its main role in the present scenarios is to serve as an abstraction layer between devices and UIs. Briefly, iCrafter is used as follows:

- Controllable entities beacon their presence by depositing self-expiring advertisements in the Event Heap. These advertisements contain a description of the service's controllable behaviors (i.e., methods and their parameters) expressed in SDL, a simple XML-based markup language we developed.
- A device capable of displaying a UI (Web browser, handheld, etc.) can make a request for the UI of a specific service, or can query the iCrafter's *Interface Manager* (which tracks

Implementation Details

Transmitting devices (buttons, sliders) contain a Ming TX-99 V3.0 300 MHz FM radio frequency (RF) transmitter and a Holtek HT-640 encoder to send 8 bits of data to a receiver board, which contains a Ming RE-99 V3.0 RF receiver and a Holtek HT-648L decoder. The receiver board sends its data to a PC using either the parallel or USB port, and a listener program running on the PC then posts an appropriate tuple (based on the ID received) to the iRoom's Event Heap. Receiving devices (buzzers, LEDs) work in the opposite manner: a listener program receives an event intended for the iStuff and sends the target device ID through either the parallel or USB port to an RF transmitter. This data is then received wirelessly by an RF receiver in the device, resulting in the desired behavior. The iSpeaker has a different architecture, since the RF technology we employed is not sufficient for handling streaming media. Instead, a listener program on the PC waits for speaker-targeted events, and in response streams sound files over an FM transmitter, which our iSpeaker (a small portable FM radio) then broadcasts. See http://stuff.stanford.edu/ and [8] for more information.

all advertisements) to request a list of available services. This initial request is made via whatever request-response technology is available on the client: visiting a well-known dynamically generated Web page is one possibility.

- The desired UI is created by feeding the SDL contained in a recent service advertisement to one or more interface generators. These may be local or remote (i.e., downloaded on demand over the Web), and may be specialized per service and/or per device. The Interface Manager determines the policy for selecting a generator. Part of this process includes integrating contextual information from a separate context database relevant to each workspace, making a "static" UI description portable across installations. For example, in a workspace such as ours with three large wall-mounted displays, it is preferable for a UI to refer to these as "Left, Center, Right" than to use generic names such as "screen0, screen1, screen2" (Fig. 2).

Note that in the last step the client device and service do not need to establish a direct connection (client-server style).This makes each robust to the failure of the other. They do not even need to be able to name each other using lower-level names such as network addresses because the tuple-matching mechanism can be based on application-level names or attributes of the service ("retrieve advertisements for all devices of type LightSwitch").The same service can be controlled from a variety of different devices without knowing in advance what types of devices are involved, since the same SDL description can be processed into quite different UIs by different interface generators.

The ability to insulate the service and UI from each other in these ways has been critical to the rapid prototyping of new UI behaviors. iStuff builds on this ability, using this indirection to enable rapid prototyping of *physical* UIs as well.

iStuff: Physical Devices for Ubiquitous Computing

iStuff Motivation and Definition

iStuff is a toolbox of wireless platform-independent physical UI components designed to leverage the iROS infrastructure (which allows our custom-designed physical devices to send and receive arbitrary commands to and from other devices, machines, and applications in the iRoom).The capability to connect a physical actuator to a software service on the fly has appeal to users of a ubiquitous computing environment such as a smart home. Residents would have the ability to flexibly set up sensors and actuators in their home, and designers of such homes would be able to prototype and test various configurations of their technology before final installation.

There are several characteristics that are crucial for our iStuff:

- Completely autonomous packaging, wireless connection to the rest of the room, and battery-powered operation
- Seamless integration of the devices with iROS as an existing, available, cross-platform ubiquitous computing environment to let devices,

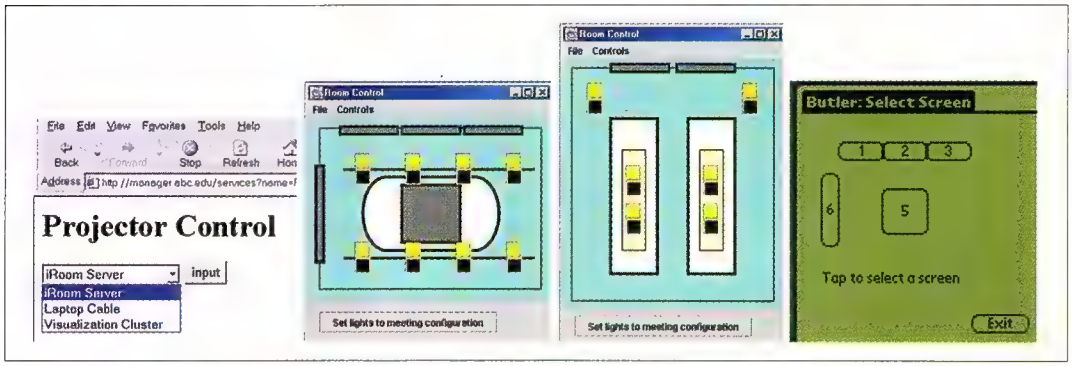


Figure 2. *Screen/projector control UIs customized for various devices and incorporating installation-specific information from the context database: (a) A fragment of a projector HTML interface that can be rendered by any Web browser. This UI was generated by a projector-specific HTML generator. The symbolic names such as iRoom Server are stored in the context database and appear in the SDL markup as "machine0," "machine1," and so on. (b) Room control applet for the same room, generated by a Java Swing-UI generator. The geometry information for drawing the widgets comes from the context database, so the generator itself is not installation-specific. Users can drag and drop Web pages onto the screen widgets to cause those documents to appear on the corresponding room screens. (c) The same UI using different geometry information (for a different room) from the context database. (d) A Palm UI rendered in MoDAL [8] that lacks the drag-and-drop feature.*

machines, and services talk to each other and pass information and control around
- Easy configuration of mappings between devices and their application functionality, by customizing application source code, or even just updating function mappings using a Web interface
- Simple and affordable circuitry

Various other research projects have looked at physical devices in the past; Ishii's Tangible Bits project [6] introduced the notion of bridging the world between bits and atoms in UIs, and more recently Greenberg's Phidgets [7] represent an advanced and novel project in physical widgets. Phidgets, however, are designed for use in isolation with a single computer, are tethered, and do not work across multiple platforms.

Device Classification and Implementation

The range of potentially useful UI components is almost unlimited, and the really useful devices to go in a standard toolbox will only be identified over time. Ideas for such devices can be categorized according to whether they are input or output devices, and the amount of information they handle, as in the following examples:
- One-bit input devices, such as pushbuttons and toggle buttons, or binary sensors such as light gates
- Multiple-bit discrete input devices, such as rotary switches or digital joysticks as well as packaged complex readers that deliver identification data as a result (e.g., barcode readers or fingerprint scanners)
- Near-continuous input devices, such as sliders, potentiometers, analog joysticks or trackballs, and various sensors (light, heat, force, location, motion, acceleration, etc.)
- Streaming input devices, such as microphones and small cameras
- One-bit output devices, such as a control or status light, beepers/buzzers, solenoids, and power switchers
- Multiple-bit discrete output devices, such as LED arrays or alphanumerical LCD displays
- Near-continuous output devices, such as servos, motors, dials, and dimmers
- Streaming output devices, such as speakers and small screens

Thus far, students in our laboratory have designed and built five types of prototype iStuff devices spanning four of the above categories: iButtons, iSliders, iBuzzers, iLEDs, and iSpeakers (Fig. 3).While our hardware designs have proven surprisingly powerful and proofs of concept, they are simple enough to be reproduced easily (see box).

We have developed several successful setups using iStuff in the Stanford iRoom:

1. New users coming into our iRoom are not familiar with the environment, and need an "entry point" to learn about the room and its features. Using our iStuff configuration Web interface, we programmed one iButton to
- Send events that turn on all the lights in the room
- Switch on all SMARTBoards (large touch-sensitive displays) and our interactive table display
- Bring up a Web-based introduction to the room on one SMARTBoard
- Show an overview of document directories for the various user groups on a second SMARTBoard
- Open up our brainstorming and sketching application on the third SMARTBoard

It is worth nothing that setting up this iRoom Starter took less than 15 minutes of configuration using the Web interface.

2. SMARTBoards provide only a rather inconvenient way to issue right clicks when using the touch-sensitive board for input — users have to press a right-click "mode key" on the tray in front of the board to have their next touch at the board be interpreted as a right click. To study whether having the right-click modifier closer to the actual input location at the board would make this interaction more fluid, we built a specialized iButton that was shaped to fit inside a hollow pen prototype made from RenShape plastic by a product design student in our model shop. When the button is pressed it sends an event to the Event Heap that is then received by a listener application running on the computer associated with the SMARTBoard. The listener then tells the SMARTBoard driver to interpret the next tap as a right click. Users can now simply press the button on the pen and then tap to issue a right click.

3. We found our iSlider could conveniently control the paddles for our multiscreen version of the classic video game Pong, described below.

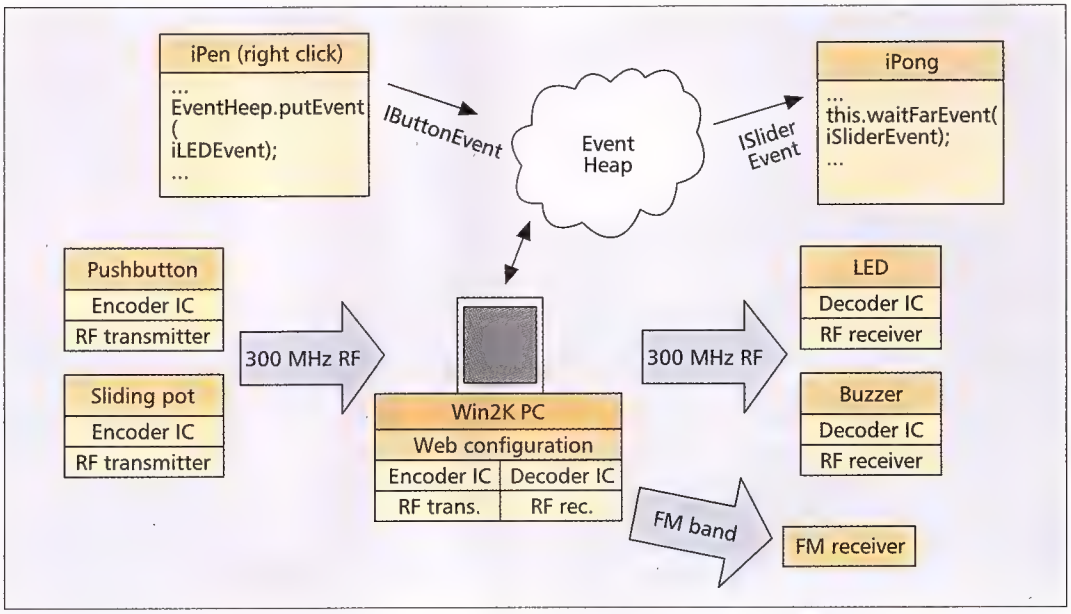
4. The iSpeaker has been extended to provide verbal feedback for user actions (e.g., "Pong game started") by means of a text-to-speech program — applications simply send a SpeakText event to the iSpeaker containing the ASCII text to be spoken.

5. We are experimenting with our iLEDs and iBuzzers to provide feedback about the status of devices in the room.

As discussed before, the Event Heap is a core component of the iROS that makes it possible to decouple the sender and receiver from the con-

Figure 3. *The various types of iStuff created so far: buttons, potentiometers, speakers, and buzzers.*

When residents acquire a new device, or wish to reconfigure existing devices, they can simply use a utility such as our "Patch Panel" to map the event type sent by the new device to the event type expected by the target application.



■ **Figure 4.** *The overall system architecture for iStuff. In darker boxes are the actual physical devices, and in the lighter ones are a couple of examples of applications using iStuff and the iROS Event Heap. An iStuff server translates the wireless device transmissions into software events, which can be consumed by interested applications. For example, when the potentiometer of the iSlider is moved, it sends a radio signal, which is received by the server and turned into a SliderEvent. The event is posted to the Event Heap, and subsequently received by the iPong application, which is listening for SliderEvents.*

tents of the message itself (Fig. 4).This architecture allows great flexibility for the prototyping of interfaces; for instance, an application can be controlled by either a traditional mouse, a graphical slider widget, or an iSlider as long as each of those devices sends the event type (perhaps an event containing a new Y-coordinate) for which the application is listening.

In our iRoom we have demonstrated the utility of the combination of Event Heap software with iStuff hardware by developing iPong, a multimachine version of the video game where players control the vertical position of virtual paddles to make contact with a virtual bouncing ball. The game was written to listen for Paddle Events, which contain information about the new position of the target paddle. Any input method that can generate a Paddle Event can control the paddle position. We have mapped the standard mouse, touch panel input, and an iSlider (a sliding-potentiometer iStuff widget) to drive the paddle. To the application, the physical source of the events is irrelevant. Thus, we have decoupled the link between hardware and software components in a physical UI.

Our iButtons are already reconfigurable dynamically via a Web interface that lets users enter arbitrary events to send when a specific button is pressed. We intend to provide this flexible interactive mechanism for mapping applications and events for all iStuff, using the on-the-fly service discovery tools of iROS (described earlier).The result will be a general virtual Patch Panel that allows even end users to map events to services and map conversions between related event types. Thus, iStuff makers can send and receive their own types of events (e.g., button events or slider events) without concern for the exact names of events desired by end-user applications, and application developers can send and receive their own types of events (e.g., Paddle Event) without prior knowledge of every possible type of device the user might choose to interface with their application.

The iStuff/Event Heap combination has direct applications to the Smart Home that incrementally acquires new technologies. When residents acquire a new device or wish to reconfigure existing devices, they can simply use a utility such as our Patch Panel to map the event type sent by the new device to the event type expected by the target application.

Smart Home Applications

While our iStuff was originally designed with our iRoom (a space used for meetings and brainstorming/design sessions) in mind, our technology and infrastructure could be useful in a smart home environment. In particular, the ability to create *task-oriented user interfaces* — interfaces reflecting the user's task as opposed to the technical features of an appliance — makes iStuff particularly compelling for Smart Home applications.

Dynamic, task-based remote controls: Currently, when a user wants to watch a movie on a DVD, they need several remote controls: one to control the DVD player, another to control their home's surround-sound system, and a third to control the television set (and then the user has to get up to dim the lights!). Today's remotes are device-based, but because the Event Heap architecture allows for the decoupling of devices from messages we are able to use iStuff to construct task-based remote controls. By gathering appropriate iStuff components and using the Patch Panel application to ensure that the appropriate iStuff events are converted to the events appropriate to the target devices

(DVD player, speakers, TV set, lights), the user can construct a task-oriented controller — one device that controls all appliances relevant to viewing a DVD movie, regardless of their physical connectivity. iCrafter could be used in an analogous manner to dynamically create GUI controllers for household appliances, thus transforming a PDA into a task-based universal remote control.

Monitoring house state: A user is on her way out the door of her smart home, about to head off to work. The display near her door shows her the status of several devices in her home that have been instrumented with iSensors: did she leave the stove on? The lights in her bedroom? Is the thermostat too high? Is the burglar alarm on?

Setting house state: A user can create an iButton or similar device to set the house's "state" as she leaves for work every day, and mount this button by her door. She might configure it to lower her thermostat, switch off all lights, and activate her security system, for example. This type of button is analogous to our Start iRoom button mentioned earlier.

Smart home design: Architects and interior designers could use iStuff to fine-tune the placement of controls, speakers, and other interactive elements of a Smart Home. Researchers and technology developers could use iStuff to quickly prototype and test their products before putting them on the market for addition to smart homes.

Discussion and Summary

Technology advancements have made much of the original vision of ubiquitous computing feasible. A software framework, however, that integrates those heterogeneous technologies in a dynamic, robust, and legacy-aware fashion and provides a seamless user experience has been missing. We have created the Stanford iRoom, a physical prototypical space for ubiquitous computing scenarios that has been in constant use for over two years now, to address this need. iROS, our iRoom Operating System, runs as a meta-OS to coordinate the various applications in the room. iROS is based on a tuplespace model, which leads to the aforementioned desired characteristics. Its failure robustness has been better than average for both induced and real faults. Its ability to leverage and extend existing applications has been critical for rapid prototyping in our research.

The iStuff project builds on iROS, and tackles the problem that customizing or prototyping physical user interfaces for ubiquitous computing scenarios (e.g., smart homes) is still a very arduous process. It offers a toolbox of wireless physical user interface components that can be combined to quickly create nonstandard user interfaces for experimentation. The iROS infrastructure has proven invaluable in making the software integration of these custom devices very straightforward. The flexibility of the technology we developed for Stanford's iRoom has potential benefits in a smart home scenario, for example, by enabling users to quickly create a customized task-based interface to a system in their home.

In all, we hope that our approach to building a software and hardware framework for ubiquitous computing environments, and the various building blocks we have implemented and deployed, are general and useful enough so that others will find them of value. For more information on our Stanford Interactive Workspaces project, and to access iStuff documentation or download iROS software, please visit our project homepage at http://iwork.stanford.edu/.

Acknowledgments

The authors would like to thank Maureen Stone, Rafael Ballagas, Robert Brydou, Michael Champlin, Hans Andersen, and Jeff Raymakers for their contributions to this work, as well as the Wallenberg Foundation (http://www.wgln.org) for its financial support.

References

[1] D. Gelernter and N. Carriero, "Coordination Languages and their Significance," *Commun. ACM*, vol. 32, no. 2, Feb. 1992.

[2] B. Johanson and A. Fox, "The Event Heap: A Coordination Infrastructure for Interactive Workspaces," *Proc. WMCSA 2002*, Callicoon, NY, June 2002.

[3] T. D. Hodes et al., "Composable Ad-Hoc Mobile Services for Universal Interaction," *Proc. ACM MobiCom '97*, Budapest, Hungary, Sept. 1997.

[4] S. R. Ponnekanti *et al.*, "ICrafter: A Service Framework for Ubiquitous Computing Environments," *Proc. UbiComp. '01*, Atlanta, GA.

[5] T. Lehman *et al.*, MoDAL (Mobile Document Application Language); http://www.almaden.ibm.com/cs/TSpaces/MoDAL

[6] H. Ishii and B. Ullmer, "Tangible bits: Towards seamless interfaces between people, bits and atoms," *Proc. CHI '97*, Atlanta, GA, Mar. 22–27, 1997, pp. 234–41.

[7] S. Greenberg and C. Fitchett, "Phidgets: Easy Development of Physical Interfaces Through Physical Widgets," *Proc. UIST 2001*, Orlando, FL, Nov. 11–14, 2001, pp. 209–18.

[8] R. Ballagas *et al.*, "iStuff: A Physical User Interface Toolkit for Ubiquitous Computing Environments," to appear, *Proc. CHI 2003*, Ft. Lauderdale, FL, Apr. 2003.

Additional Reading

[1] W. K. Edwards and R. E. Grinter, "At Home with Ubiquitous Computing: Seven Challenges," *Proc. UbiComp '01*, Atlanta, GA, pp. 256–72.

[2] E. Kiciman and A. Fox, "Using Dynamic Mediation to Integrate COTS Entities in a Ubiquitous Computing Environment," *Proc. HUC2K*, LNCS, Springer Verlag.

Biographies

JAN BORCHERS (borchers@cs.stanford.edu) is an acting assistant professor of computer science at Stanford University. He works on human-computer interaction in the Stanford Interactivity Lab, where he studies post-desktop user interfaces, HCI design patterns, and new interaction metaphors for music and other types of multimedia. He holds a Ph.D. in computer science from Darmstadt University, and has been known to turn his research into public interactive exhibits.

MEREDITH RINGEL (merrie@cs.stanford.edu) is a second-year Ph.D. student in computer science at Stanford, with a focus on human-computer interaction. She received her B.S. in computer science from Brown University.

JOSHUA TYLER (jtyler@cs.stanford.edu) is a researcher at HP Labs, Palo Alto, California. He holds a Master's degree in computer science from Stanford with a specialization in human-computer interaction, and a B.S. in computer science from Washington University.

ARMANDO FOX (fox@cs.stanford.edu) joined the Stanford faculty in January 1999. His research interests include the design of robust Internet-scale software infrastructure, particularly as it relates to the support of mobile and ubiquitous computing, and user interface issues related to mobile and ubiquitous computing. He received a B.S.E.E. from M.I.T., an M.S.E.E. from the University of Illinois, and a Ph.D. from UC Berkeley. He is a founder of ProxiNet, Inc. (now a division of PumaTech), which commercialized thin client mobile computing technology developed at UC Berkeley.

A user can create an iButton or similar device to set the house's "state" as she leaves for work everyday, and mount this button by her door. She might configure it to lower her thermostat, switch off all lights, and activate her security system, for example.

FACILITATING THE PROGRAMMING OF THE SMART HOME

JENS H. JAHNKE, MARC D'ENTREMONT, AND JOCHEN STIER, UNIVERSITY OF VICTORIA

The ongoing miniaturization and cost reduction in electronic hardware has created opportunity for equipping homes with inexpensive smart devices for controlling and automating various tasks in our daily lives. Networking technology and standards have an important role in driving this development.

ABSTRACT

The ongoing miniaturization and cost reduction in the sector of electronic hardware has created ample opportunity for equipping private households with inexpensive smart devices for controlling and automating various tasks in our daily lives. Networking technology and standards have an important role in driving this development. The omnipresence of the Internet via phone lines, TV cable, power lines, and wireless channels facilitates ubiquitous networks of smart devices that will significantly change the way we interact with home appliances. Home networking is foreseen to become one of the fastest growing markets in the area of information technology. However, interoperability and flexibility of embedded devices are key challenges for making smart home technology accessible to a broad audience. In particular, the software programs that determine the behavior of the smart home must facilitate customizability and extensibility. Unlike industrial applications, typically engineered by highly skilled programmers, control and automation programs for the smart home should be understandable by laypeople. In this article we discuss how recent technological progress in the areas of visual programming languages, component software, and connection-based programming can be applied to programming the smart home. As an example of an industrial prototype solution, we present microCommander, a visual tool for rapidly programming synergetic devices for the smart home.

PROGRAMMING CHALLENGES FOR THE SMART HOME

The ongoing miniaturization and cost reduction in the sector of electronic hardware has created ample opportunity to equip private households with inexpensive smart devices for controlling and automating various tasks in our daily lives. Networking technology and standards play an important role in driving this development. The omnipresence of the Internet via phone lines, TV cable, power lines, and wireless channels facilitates ubiquitous networks of smart devices that will significantly change the way we interact with home appliances. Home networking is expected to become one of the fastest growing markets in the area of information technology. However, interoperability and flexibility of embedded devices are key challenges for making *smart home* technology accessible to a broad audience. An increasing number of connectivity standards for net-centric smart devices have been proposed by companies and industrial consortia, such as *HAVi* (home audio-video interoperability), *JetSend* (intelligent service negotiation), *Jini*, and *Bluetooth* (proximity-based wireless networking) [1]. In particular, Bluetooth has gained a lot of attention and momentum in industry. A large number of companies led by Ericsson, Nokia, IBM, Toshiba, Intel, 3Com, Motorola, Lucent Technologies, and Microsoft have released or announced Bluetooth-enabled products, most of them in the areas of home entertainment, mobile telecommunication and personal information management.

Still, connectivity standards solve only the first part of the integration problem, dealing with the creation of a common *channel* for communicating among various smart appliances. The second part of the problem is to establish a common *language* so that home appliances can actually understand each other and function in a collaborative manner. In general, this problem of *semantic interoperability* is much harder to solve than the realization of the physical transport channel for data. The main reason for these difficulties is the great *heterogeneity* of home appliances and the large variety of their embedding context. Home appliances cover all aspects of our daily lives including environmental controls, lighting, alarm systems and security, telecommunication, cooking, cleaning, and entertainment. There exist a vast number of potential scenarios for integrating such appliances. It is not possible for vendors to foresee all these applications and equip their devices with functionality that enables collaboration with every other device a customer would like to integrate. Consequently, there is the need for customization mechanisms that can be used for integrating different appliances and sensors into a common process that controls the smart home.

Such customization mechanisms can be seen as the "programming language" for the smart home. Primary requirements for such a programming language are *ease of use* and *rapid deployment*. Unlike industrial applications that

1536-1284/02/$17.00 © 2002 IEEE

are typically engineered by highly skilled programmers, control and automation programs for the smart home should be understandable by laypeople. Analogous to other "do-it-yourself" maintenance activities around the home, programs for the smart home should be changeable by third-party service providers as well as the homeowner herself. There is a good chance of achieving this goal because applications in home automation tend to have lower complexity than industrial automation systems. Still, traditional programming paradigms like textual programming languages appear inadequate for this purpose. Effective programming mechanisms for the smart home require innovative paradigms that lift programming to a level of abstraction that is similar to plugging in a new stereo or TV set. We will introduce three such innovative paradigms in the following section. Then we will present an example solution for programming the smart home.

Enabling Paradigms

Driven by the rapidly increasing complexity of software applications over the last decades, the area of *software engineering* has been established as a discipline of systematic construction and maintenance of quality software systems. Software engineering encompasses a broad spectrum of aspects including all life cycle activities starting from the requirements analysis of a new system up to the reverse engineering and modernization of outdated legacy systems. In this article we discuss three emerging software engineering paradigms that, in combination, have great potential for facilitating the programming of the smart home. These paradigms are *visual programming*, *component-based software construction*, and *connection-based programming*. The following three subsections introduce the basic ideas behind these three paradigms. Then we use an application example to illustrate how concerted use of these paradigms can facilitate the programming of home automation systems.

Visual Programming Languages

The development of visual programming languages (VL) has been driven by the experience that developers tend to understand pictures better than plain program text. One reason for their increased expressiveness is that pictures have a two-dimensional nature in contrast to sequential program text, which covers only one dimension. Visual formalisms have been used extensively as an aid to design and visualize algorithms including their control flow or data flow. Classic examples of such formalisms are flow charts or Nassi-Schneiderman diagrams. More recently, similar concepts have been adopted in the Unified Modeling Language (UML) in the form of activity diagrams (www.uml.org). Flow charts of this kind can easily be mapped to equivalent constructs in textual programming languages. Several software engineering tool vendors offer development environments that can perform this mapping automatically, making textual programming superfluous for applications of low or medium complexity. Today, visual programming languages are often used in combination with textual languages. Moreover, visual languages and software visualization paradigms are increasingly used to aid human understanding of the existing program code in legacy systems.

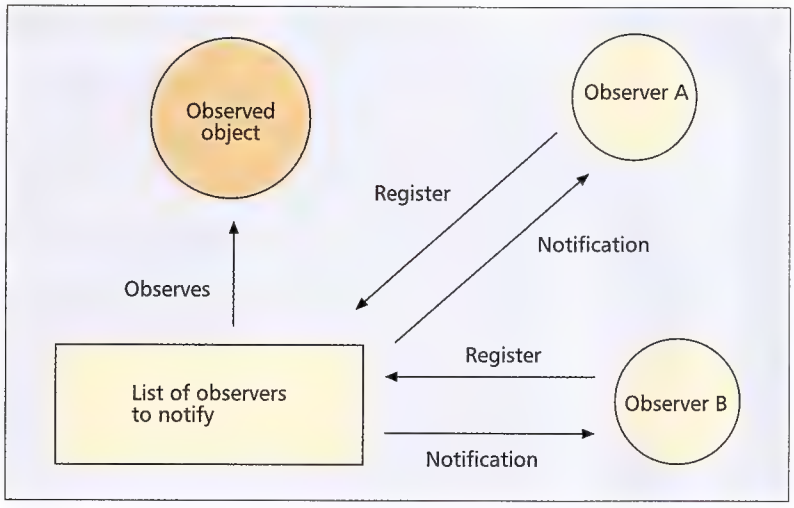
Apart from considerations of program *logic*, the area of visual languages was equally driven by progress in the domain of user interface design and human-computer interaction. Several so-called fourth-generation languages (4GL) introduced in the '80s and '90s included visual tools for user interface design as an integral component. This paradigm has been broadly adopted with popular programming tools like Microsoft's Visual Basic or, more recently, IBM's VisualAge for Java. Such visual programming languages typically promote event-driven architectures. This means that the programmer does not explicitly define the control flow; it is implicitly determined by the occurrence of user interface events (e.g., a mouse click on a button). Both visual programming paradigms, flow logic diagrams and event-driven user interface designs, are complementary rather than competing approaches. They can be integrated into a holistic solution for visual programming. We give an example of such a solution later in this article.

Component Software

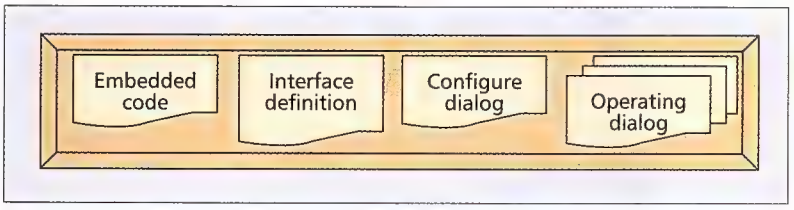
The idea of component software has its roots in the great success of component-based manufacturing in the hardware sector. Component-based software systems are assembled from a number of preexisting pieces of software called *software components* (plus additional custom-made program code). Software components should be (re)usable in many different application contexts. Particularly, these components should be usable in unpredicted applications and by third parties. The term commercial off-the-shelf (COTS) was coined in the mid '90s as a concept for a binary piece of commercial software with a well defined application programming interface and documentation. The component market has gained momentum from the introduction of infrastructure for deploying components in programming languages and operating systems, such as Sun Microsystem's (Enterprise) Java Beans and Microsoft's COM+. Using the component paradigm for software construction has various benefits: it increases the degree of abstraction during programming, provides proven (error-free) solutions for certain aspects of the application domain, increases productivity, and facilitates maintenance and evolution of software systems.

Component-based software development has become an important part of modern software engineering methods. So-called *lightweight* components (i.e., fairly small in size) are already well understood. Lightweight components have become part of modern programming languages (e.g., the Swing library within Java). Most modern approaches to user interface development are component-based. In contrast to lightweight components in particular application domains, the construction and reuse of more heavyweight components in arbitrary application domains still pose many theoretical questions. Examples of such research questions are: how can a complex

Both visual programming paradigms, flow-logic diagrams and event-driven user interface designs, are complementary rather than competing approaches. They can be integrated into a holistic solution for visual programming.



■ Figure 1. *The observer pattern as the basis for connection-based programming.*



■ Figure 2. *Aspects of the microCommander component.*

component be documented so that it can be well understood by a third party? How can a developer find the components that best fit her current software project?

Still, our research indicates that the current state of component technology in software engineering is sufficient as an enabling paradigm for constructing software for home automation. In other words, we believe that libraries of relatively lightweight components are adequate to represent the building blocks to construct programs for the smart home. We present an example of such a library later.

Connection-Based Programming

Rather than being an independent paradigm in its own right, the introduction of connection-based programming has been driven mainly by the introduction of the previously discussed paradigms, component software and visual languages. Traditional software programs have followed the procedure call paradigm, where the procedure is the central abstraction called by a client to accomplish a specific service. Programming in this paradigm requires that the client has intimate knowledge about the procedures (services) provided by the server. However, this kind of knowledge is not present in component software because it is based on components from third parties that were separately developed. That is why component software requires a new programming paradigm called *connection-based programming*.

In connection-based programming, connections between pieces of software are not implicitly defined by procedure calls but are explicitly programmed. Connections represent the glue that binds together interfaces of different software components. The basis for connection-based programming is typically the so-called *observer* design pattern, which is shown in its general form in Fig. 1. The observer pattern works like a subscription mechanism that handles callbacks upon the occurrence of events. Software components interested in an event that could occur in another component can register a callback procedure with this component. This procedure is called every time the event of interest occurs. The interfaces of software components have to be tailored for connection-based programming — they have to provide subscription functions for all internal events that might be of interest to external components. This part of the interface is often called the *outgoing* interface of a component, as opposed to its *incoming* interface that consists of all callable service procedures.

Bringing It Together for Programming the Smart Home

In this section we discuss how the three enabling paradigms introduced in the previous section can be used to facilitate programming of the Smart Home. We report on the results of an industrial-driven collaborative research project carried out between the University of Victoria, British Columbia, Canada, and Intec Automation Inc., a Victoria company in the area of embedded systems. The project is supported by the Advanced Systems Institute of British Columbia.

Embedded Programming with Visual Components

Over the last few years, Intec has investigated how the visual programming paradigm can be used to facilitate the development of embedded control applications for industrial as well as private applications. As a result, Intec has developed *microCommander*, an application for presenting a visual programming interface to a system of embedded devices (www.microcommander.com). The software runs on a personal computer with access to a network connecting any number of devices. All of the devices must conform to a predefined component architecture that is recognized by microCommander. MicroCommander then allows a user to visually program these off-the-shelf software components without having to write any source code. Behind the scenes, each component consists of an *embedded code*, an *interface and behavior definition*, a *configure dialog*, and *operating dialogs* (Fig. 2). The *embedded code* containing the logic that operates a device usually executes on microcontrollers located at various places within the automated home. Depending on the complexity of the device, a single microcontroller may host the *embedded code* for a number of components. The *interface and behavior definition* describes how to interact with the device in terms of input and output messages. Messaging formats and policies are part of the component architecture, and it is critical that every device in the system strictly conform to these rules. Thus, it is guaranteed that every device is addressable and properly controllable.

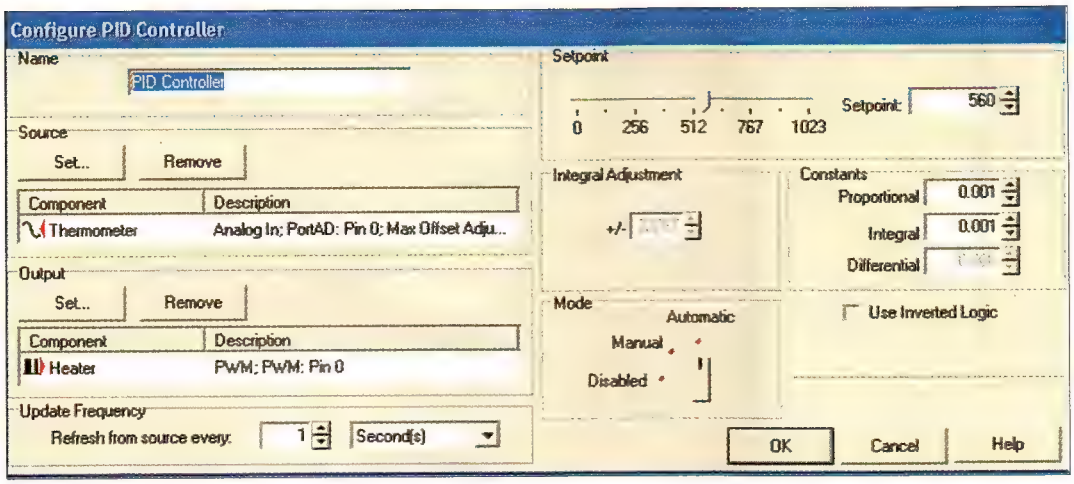


Figure 3. *The configure dialog allows specification of a number of component-related parameters.*

The *configure dialog* (Fig. 3) is a visual interface for programming properties such as microcontroller input and output assignments, default values, and states. This configuration setup is generally done only once during system installation, after which the component is exclusively controlled via the *operating dialogs*. The use of *configure dialogs* requires some domain knowledge, and thus would typically be done by third-party vendors during installation. For example, Fig. 3 shows the setup of a new heater system controlled by a PID control component [2].

The *operating dialogs* (Fig. 4) provide sophisticated visual interfaces to the devices embedded within the home. Each *operating dialog* is customized for the day-to-day usage of a device by a layperson. Both *configure dialogs* and *operating dialogs* reside within the microCommader application, and are part of its user interface. The application contains an extensible library of visual controls that the operating dialogs may utilize. MicroCommander thus acts as a PC-based remote control console to the device, allowing a home owner to manipulate and visually program a home from any Internet-ready PC running microCommander.

During initialization, microCommander learns about the devices present in the network by a process called *introspection*, in which each device automatically identifies itself. Operating platforms such as the Bluetooth wireless connectivity protocol or the Java Jini framework already provide mechanisms for introspection that support addressing, security, and authentication [1]. During introspection, each device publishes an *operating dialog* and, depending on security parameters, a *configure dialog*. At regular time intervals introspection is repeated in order to keep the view of the system up to date. Devices that were previously turned off or recently installed are automatically discovered.

The *operating dialogs* generally resemble real-world artifacts users are accustomed to (Fig. 4). Appliances appear on the application canvas as

In connection-based programming, connections between pieces of software are not implicitly defined by procedure calls but they are explicitly programmed. Connections represent the glue that binds together interfaces of different software components.

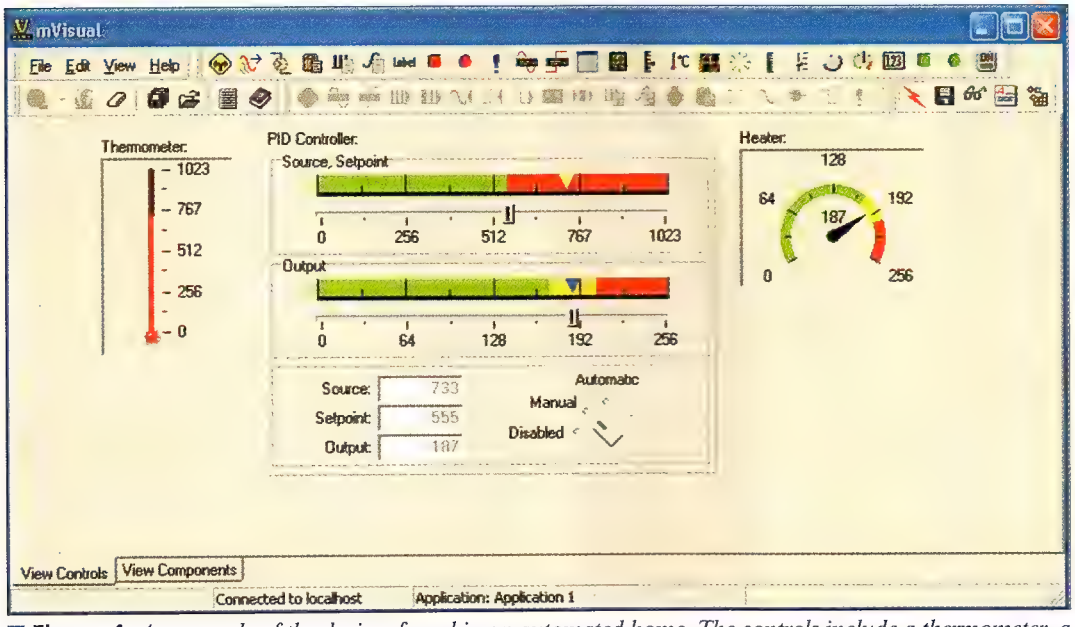


Figure 4. *An example of the devices found in an automated home. The controls include a thermometer, a furnace temperature gauge, and a heater controller.*

The microSynergy framework presents a development and execution environment for the connection-based programming paradigm. As a result of this programming paradigm the embedded appliances are unaware of the fact that they are part of a collaborative network.

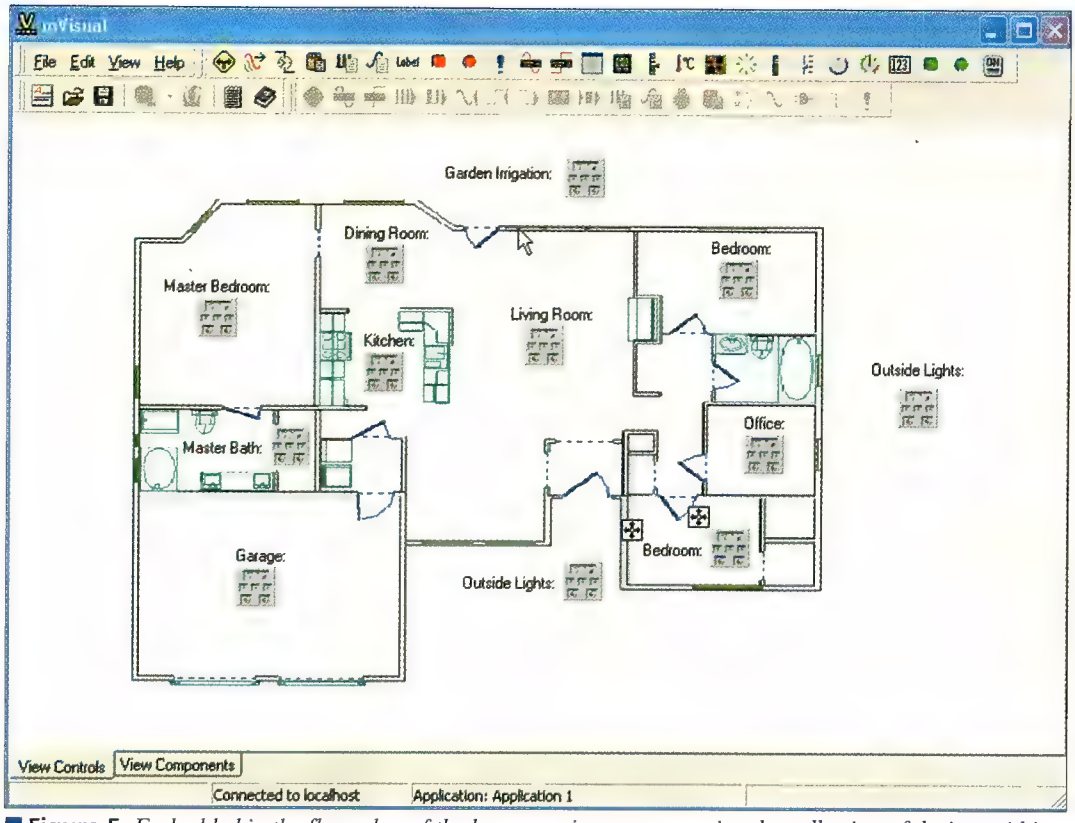


Figure 5. *Embedded in the floor plan of the house are icons representing the collection of devices within each room. By clicking on an icon a user gets a detailed view of the devices within that room (Fig. 4).*

familiar images that represent the appliances' current state by means such as plain text or colored gauges. Each component type may publish a range of *operating dialogs* that either provide a different look and feel, or are tailored to varying types of display devices. Input values to devices are changed via their graphical representation, by either entering new values in text fields or moving pictures of levers, switches, and gauges using the mouse. The *operating dialog* first translates this type of activity into commands understood by the corresponding software component, and then transmits that information in real time to the embedded code. The entire process is similar to walking up to a device and physically turning a knob. However, here it is done via a computer screen and mouse from any Internet-ready PC running the microCommander application.

As a result of periodically querying each component's state, the *operating dialog* provides feedback to changes as they occur in real time. For example, the turning of a light switch may immediately become apparent in a light sensor changing state, and adjusting the temperature of the furnace may induce a slow continuous change in a thermometer reading.

By filtering, grouping, and rearranging controls on the screen, different views of the automated home are possible (Fig. 5). Devices may be grouped by type, location, or relevance, allowing the creation of interfaces that are optimized for particular tasks such as power consumption, home security, or temperature control. Background images such as floor plans, wiring, and plumbing diagrams can be embedded into the view to provide a context in which to place devices.

Applications like microCommander act as visual control consoles for all the electronic devices embedded in the smart home, similar to the terminals used in cockpit or bridge controls of modern airliners and cruise ships. The underlying complexity of the system implementation is hidden, and the interface is reduced to the essentials. Moreover, the microCommander framework provides a general platform for visual component-based software assembly, without the need for textual programming languages. The typical edit-compile-test cycle of traditional software programming does not exist. Instead, a system is assembled and configured in real time using visual off-the-shelf components. Third-party vendors can easily develop their own net-enabled devices for use with the microCommander application by providing a conforming component interface to the device.

Using Visual Flow Charts for Connecting Smart Devices

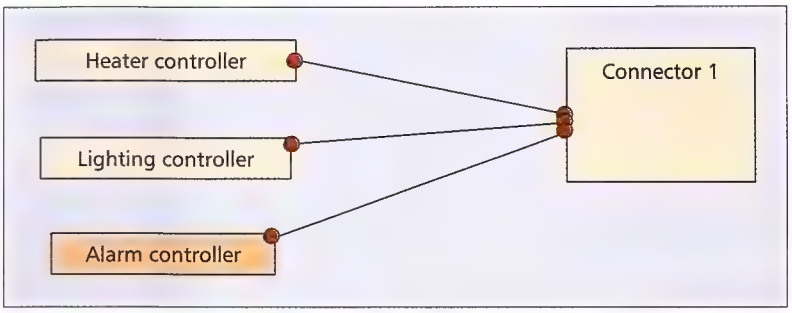
So far, we have discussed how visual programming is used for programming devices in the smart home. We now look at connecting many such devices into an integrated collaborative network. For this purpose, the University of Victoria and Intec Automation have jointly developed a technology prototype called microSynergy. MicroSynergy facilitates the development and execution of logic described with the Specification and Description Language (SDL). SDL has unambiguous formal semantics, is well accepted by the embedded systems engineering domain, and is easily understood by a layperson.

MicroSynergy consists of a *microSynergy editor* and a *microSynergy runtime engine*. Specifications created using the editor are downloaded to the runtime engine, which then controls the corresponding embedded devices accordingly. The microSynergy application applies *introspection* to discover the (type of) devices present within the home. This introspection mechanism is realized by sending out generic query messages to all possible device addresses when requested by the user of the microSynergy editor. As a result, the microSynergy editor is aware of every device in the system as well as the types of messages a device produces and the types of messages it understands. Similar to microCommander, the microSynergy editor also presents a user interface to the embedded device. However, in this case the interface omits the configure dialogs and operating dialogs and instead presents a view of the *interface definition*. Control logic is implemented in terms of input and output messages rather than visual controls. Editing SDL diagrams and establishing connections among the entities on the screen is done by simply clicking and dragging objects using the mouse and keyboard.

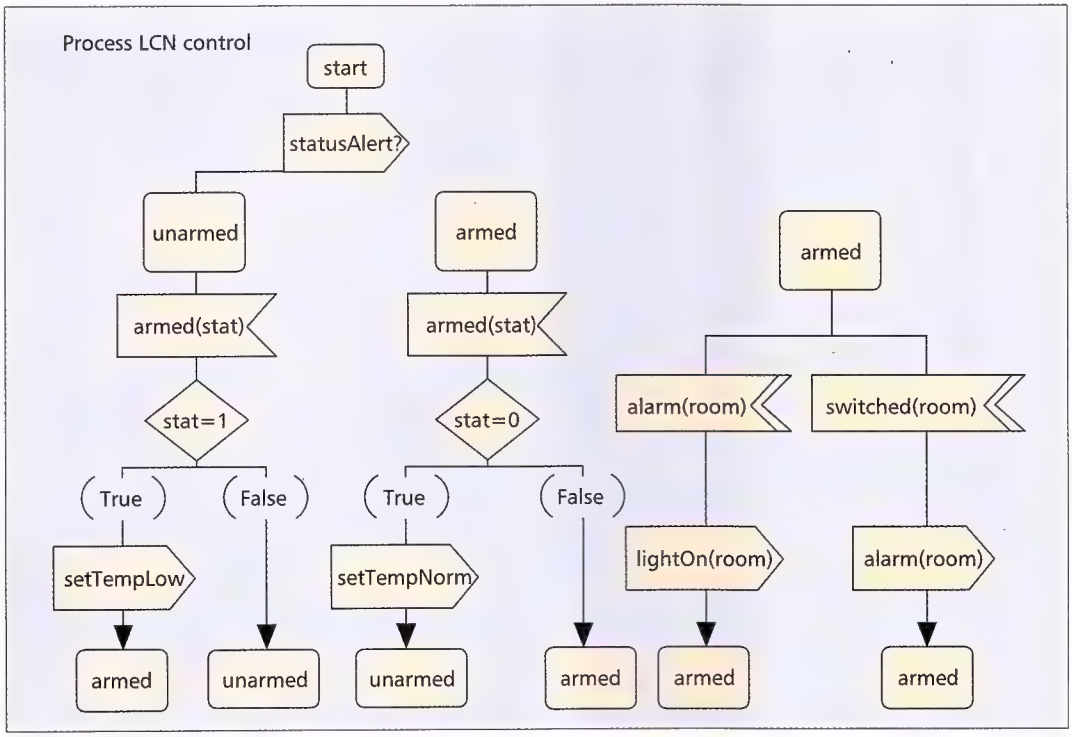
A home alarm system is an example of the need to establish elaborate logical dependencies in such a way that the events triggered by one device cause a response in another. The home alarm system, for example, once activated ought to trigger the lighting and video surveillance system. It should also automatically deactivate these systems once the threat to the home has passed. This type of scenario is easily programmed using microSynergy and SDL.

The first programming step is to connect the appropriate devices to an SDL block (Fig. 6). This step creates exclusive communication channels among the devices, ensuring that the inputs and outputs incorporated in the subsequent SDL logic are authentic. Devices may be connected to many different blocks and hence partake in separate activities. Within each SDL block resides a detailed specification of the dependencies among the connected devices (Fig. 7). Our example shows an SDL block that connects the heater controller with the alarm system and lighting control. The specification consists of *states* (represented as rectangles with round corners), *transitions* (represented by directed arcs), and conditions (diamond-shaped boxes). The flow of logic occurs from state to state via transitions. Conditional statements provide a selection mechanism for the next states. Communication with the corresponding devices occurs whenever a transition is traversed. One type of transition generates an output signal that is sent to the corresponding device, while another type of transition waits until an input signal is received. For example, the input signal at transition *armed(stat)* subsequently triggers the transition *setTempLow* if the value of *stat* is true.

After an SDL specification is complete, it can be downloaded to an embedded controller that



■ Figure 6. *A logical connector connected to three controllers.*



■ **Figure 7.** *The creation of internal connector logic with microSynergy.*

Future generations of cell phones may use Java for enabling more sophisticated interaction with the Smart Home. Moreover, emerging net-centric standards for service registry can be used to integrate the smart devices within a smart home with other community services.

executes the microSynergy runtime engine. The runtime engine provides a parallel execution platform for any number of SDL blocks. Although execution within each block is sequential, several different blocks can run at the same time.

The microSynergy framework presents a development and execution environment for the connection-based programming paradigm. As a result of this programming paradigm the embedded appliances are unaware of the fact that they are part of a collaborative network. New devices are easily added to the system, and the resulting new dependencies are quickly created.

Future Perspectives

If at least one of the devices embedded in the smart home supports TCP/IP, a small Web server can be deployed so that services may be offered via Simple Object Access Protocol (SOAP), an emerging standard for net-centric interfaces of all kinds. It combines the simple HTTP protocol with the universal data description language XML. With this kind of gateway, homeowners can use devices like Web-enabled cell phones to interact with the network just as easily as a Web browser. Future generations of cell phones may use Java to enable more sophisticated interaction with the smart home. Moreover, emerging net-centric standards for service registry like *Universal Description, Discovery, and Integration* (UDDI) and the newly standardized *Web Service Description Language* (WSDL) can be used to integrate the smart devices within a smart home with other community services. In this context, the capability of a connector being treated as a component enables scalability with respect to nested network communities. It can provide for hierarchies of components and therefore hierarchies of network communities. We will continue our collaborative research efforts in this direction.

Acknowledgments

We would like to thank Intec Automation for their collaboration on this research project. Furthermore, we thank the Advanced Systems Institute of British Columbia for supporting the research. Finally, thanks to Andrew McNair for his support in implementing the microSynergy editor.

References

[1] Sun Microsystems, "Jini Technology and Emerging Network Technologies," 2001; http://www.sun.com/jini/whitepapers/technologies.html

[2] Goodwin, Graebe, and Salgado, *Control System Design*, Prentice Hall, 2000.

Additional Reading

[1] Mitchele-Thiel, *Systems Engineering with SDL*, Wiley, 1997.

Biographies

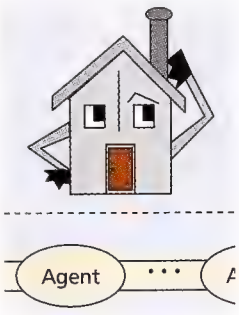
JENS JAHNKE (jens@cs.uvic.ca) is an assistant professor in the Department of Computer Science at the University of Victoria, Canada. He holds a Ph.D. degree from the University of Paderborn, Germany (1999) and an M.Sc. degree from the University of Dortmund, Germany (1994). He is a Fellow of the Advanced Systems Institute of British Columbia and was awarded the prestigious Ernst-Denert Software Engineering Award (2000). His current research interest lies in advanced topics of network-centric software engineering.

MARC D'ENTREMONT (mdentrem@cs.uvic.ca) is an M.Sc. candidate in the Department of Computer Science at the University of Victoria, Canada. He holds a B.A. in economics (1993) from the University of Guelph, and a B.Sc. (2002) from the University of Victoria, Canada. His research interests involve visual languages and evolution of embedded networks.

JOCHEN STIER (jstier@cs.uvic.ca) is a Ph.D. candidate in the Department of Computer Science at the University of Victoria, Canada. He holds an M.Sc. (1996) and a B.Sc. (1994) from the University of Victoria. His research interests are in the area of rapid development techniques for mechatronic systems. Prior to his current position, he held various positions as a software engineer.

THE ROLE OF PREDICTION ALGORITHMS IN THE MAVHOME SMART HOME ARCHITECTURE

SAJAL K. DAS, DIANE J. COOK, AMIYA BHATTACHARYA, EDWIN O. HEIERMAN III, AND TZE-YUN LIN, UNIVERSITY OF TEXAS AT ARLINGTON



The goal of the MavHome project is to create a home that acts as a rational agent. The agent seeks to maximize inhabitant comfort and minimize operation cost. To achieve these goals, the agent must be able to predict the mobility patterns and device usages of the inhabitants.

This work was supported by National Science Foundation ITR grant IIS-0121297.

ABSTRACT

The goal of the Managing an Adaptive Versatile Home (MavHome) project is to create a home that acts as a rational agent. The agent seeks to maximize inhabitant comfort and minimize operation cost. In order to achieve these goals, the agent must be able to predict the mobility patterns and device usages of the inhabitants. In this article, we introduce the MavHome project and its underlying architecture. The role of prediction algorithms within the architecture is discussed, and three prediction algorithms that are central to home operations are presented. We demonstrate the effectiveness of these algorithms on synthetic and/or actual smart home data.

INTRODUCTION

The **MavHome** smart home project is a multidisciplinary research project at the University of Texas at Arlington focused on the creation of an intelligent and versatile home environment. Our goal is to create a home that acts as a rational agent, perceiving the state of the home through sensors and acting on the environment through effectors (in this case, device controllers). The agent acts in a way to maximize its goal, which is a function that maximizes comfort and productivity of its inhabitants and minimizes operation cost. In order to achieve these goals, the house must be able to reason about and adapt to its inhabitants. In particular, a smart home agent must be able to accurately predict mobility and other activities of its inhabitants. Using these predictions, the home can accurately route messages and multimedia information, and can automate activities that would otherwise be manually performed by the inhabitants.

MavHome operations can be characterized by the following scenario. At 6:45 a.m., MavHome turns up the heat because it has learned that the home needs 15 minutes to warm to optimal temperature for waking. The alarm goes off at 7:00, which signals the bedroom light to go on as well as the coffee maker in the kitchen. Bob steps into the bathroom and turns on the light. MavHome records this interaction, displays the morning news on the bathroom video screen, and turns on the shower. While Bob is shaving, MavHome senses that Bob is two pounds over his ideal weight and adjusts Bob's suggested menu. When Bob finishes grooming, the bathroom light turns off while the kitchen light and menu/schedule display turns on, and the news program moves to the kitchen screen. During breakfast, Bob notices that the floor is dirty and requests the janitor robot to clean the house. When Bob leaves for work, MavHome secures the home, and starts the lawn sprinklers despite knowing the 70 percent predicted chance of rain. Later that day, MavHome places a grocery order for milk and cheese. When Bob arrives home, his grocery order has arrived and the hot tub is ready.

A number of capabilities are required for this scenario to occur, including data collection, activity prediction, wireless communication between multiple cooperating agents, and multimedia technologies. All of these capabilities must be seamlessly connected in a modular architecture. In this article we present such an architecture that supports the MavHome smart home project, and then focus on several prediction capabilities that are needed to realize the types of activities described above.

In particular, machine learning techniques are required to predict inhabitant movement patterns, tasks, and typical interactions with the house, and to use that information in automating house decisions, routing information, and optimizing inhabitant comfort, security, and productivity. This article introduces three such algorithms. The first predicts the mobility of the inhabitant using ideas from information theory. The second builds on a sequence matching algorithm to predict inhabitant interactions with the smart home, and the third identifies significant patterns of inhabitant activity that could be automated by the home. We validate these algorithms on synthetic data sets as well as on activity data collected from real device usage histories.

MAVHOME ARCHITECTURE

The MavHome architecture is a hierarchy of rational agents that cooperate to meet the goals of the overall home. Figure 1 shows the architecture of a MavHome agent. The technologies within each agent are separated into four cooperating layers. The **decision** layer selects actions

1536-1284/02/$17.00 © 2002 IEEE

MavHome is unique in combining technologies from databases, artificial intelligence, mobile computing, robotics, and multimedia computing to create an entire smart home that acts as a rational agent.

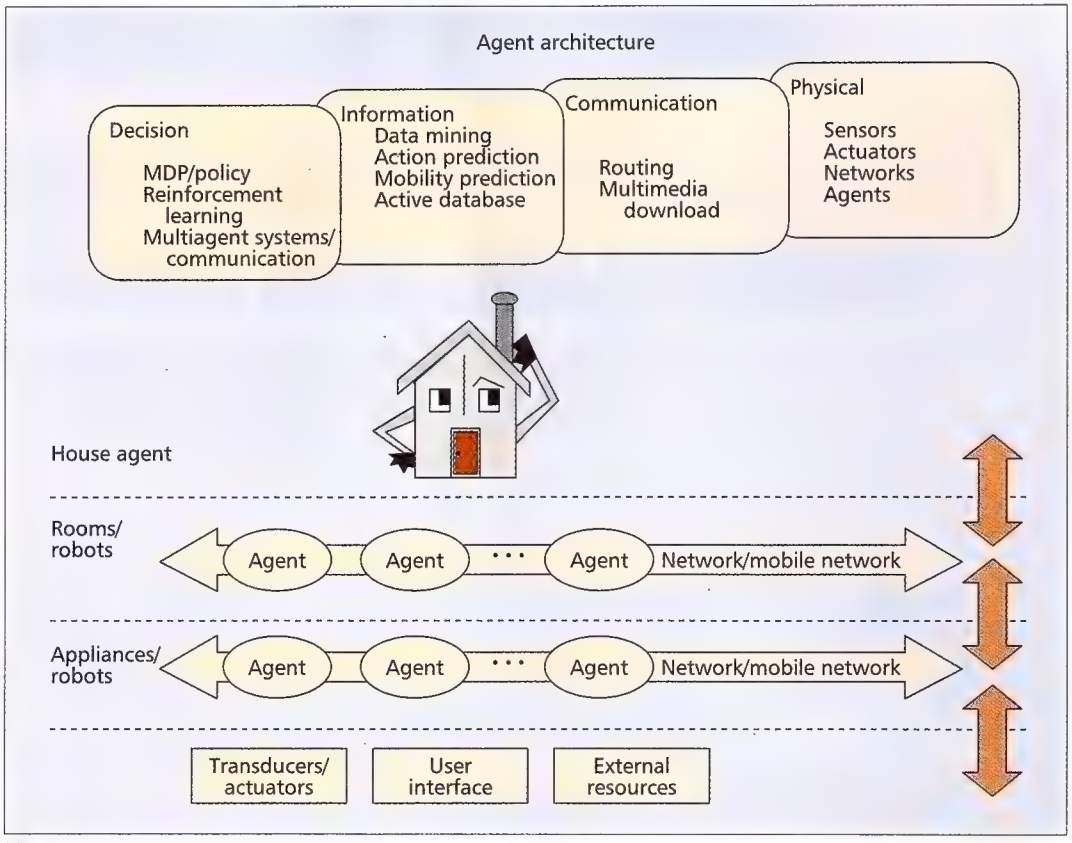


Figure 1. *MavHome agent architecture.*

for the agent to execute based on information supplied from other layers. The **information** layer gathers, stores, and generates knowledge useful for decision making. The **communication** layer includes software to format and route information between agents, between users and the house, and between the house and external resources. The **physical** layer contains the basic hardware within the house including individual devices, transducers, and network hardware. Because the architecture is hierarchical, the physical layer may actually represent another agent in the hierarchy.

Perception is a bottom-up process. Sensors monitor the environment (e.g., lawn moisture level) and, if necessary, transmit the information to another agent through the communication layer. The database records the information in the information layer, updates its learned concepts and predictions accordingly, and alerts the decision layer of the presence of new data. During action execution, information flows top down. The decision layer selects an action (e.g., run the sprinklers) and relates the decision to the information layer. After updating the database, the communication layer routes the action to the appropriate effector to execute. If the effector is actually another agent, the agent receives the command through its effector as perceived information and must decide the best method of executing the desired action. A specialized interface agent provides interaction capabilities with users and with external resources such as the Internet. As shown in Fig. 1, agents can communicate with parent/child agents or with other agents at the same level in the hierarchy.

Several smart home-related projects have been initiated by research labs. The Georgia Tech Aware Home and MIT Intelligent Room include an impressive array of sensors to determine user locations and activities within an actual house. The Neural Network House at the University of Colorado Boulder employs a neural network to control heating, lighting, ventilation, and water temperature in a manner that minimizes operating cost. The interest of industrial labs in smart home and networked appliance technologies is evidenced by the creation of Jini, Bluetooth, and Session Initiation Protocol (SIP) standards, and by supporting technologies such as Xerox PARC's Zombie Board, Microsoft's Easy Living project, the Cisco Internet Home, and the Verizon Connected Family project. MavHome is unique in combining technologies from databases, artificial intelligence, mobile computing, robotics, and multimedia computing to create an entire smart home that acts as a rational agent.

Location Prediction

Seamless connectivity is an absolute necessity for designing an intelligent environment such as MavHome. To satisfy these connectivity requirements, the smart home needs to track down an inhabitant both inside and within surrounding areas. This is the primary scope of the *location management* problem in a smart home, which exhibits some similarities with that of a typical wireless infrastructure network such as the land-mobile phone system. Wireless terminals are usually integrated in the sensors deployed in a

smart home environment and are to be worn by the inhabitants.

The MavHome coverage area is partitioned into *zones* or *sectors*. Location management involves two types of activities. When MavHome needs to contact an inhabitant, the system initiates a search for the target terminal device by polling all zones where it can possibly be found. All terminals listen to the broadcast page message, and only the target sends a response. The paging process becomes particularly inefficient if a large number of zones cover the entire smart environment. Unfortunately, due to inherent restrictions of the sensor technology such as infrared, these situations are sometimes unavoidable.

To control the location uncertainty of the inhabitant, MavHome must rely on time-to-time location update by the terminal device. This technique limits the search space for the next paging at the cost of a few location updates. Most paging algorithms can place high reliance on the latest update information, using the last known position and its surroundings as the most probable current positions [1–3]. Had probabilistic profile information been available for the inhabitant, it is also possible to modify the search space accordingly as proposed for some cellular phone systems [4].

We take a novel adaptive approach to the location management problem that is optimal in terms of both update and paging costs. The objective of our update scheme is to learn user mobility, endowing the paging mechanism with a predictive power that reduces average paging cost. Unlike earlier schemes, this provides an online algorithm for optimizing the paging problem, and allows both the paging process and the update mechanism to be considered. Machine learning methods are used to automate this process.

We represent a MavHome network (along with its neighborhood environment) by a bounded-degree connected graph $G = (\vartheta, \varepsilon)$, where node set ϑ represents the zones and edge set ε represents the neighborhood (walls, hallways, etc.) between pairs of zones. Figure 2 shows a home floor plan with 15 zones and the corresponding graph representation.

Movement History and the Mobility Model

The real power of an adaptive algorithm comes from its ability to learn. As events unfold, a learning system observes the history to use it as a source of information. Referring to Bob's scenario, let us now look at Bob's movement history for one day within the house as shown in Fig. 2. Table 1 shows the events that make him cross zone boundaries. Let us deploy a pure movement-based scheme that generates an update whenever a zone boundary crossing is detected. The system thus receives a sequence of zone ids as input yielding the movement history *mamcmrkdkdgoogdklrmcmamrlkdlrmamc*....

The *movement history* of a user is represented by a string "$\upsilon_1\upsilon_2\upsilon_3$..." of symbols from the alphabet ϑ, where ϑ is the set of zones in the house and υ_i denotes the zone id reported by the ith update.

In contrast to a movement history, a mobility model is probabilistic and extends to the future. The tacit assumption is that an inhabitant's movement (favorite routes and habitual duration of stay) is merely a reflection of the patterns of his/her life, and those can be learned. Learning aids decision making when reappearance of those patterns is detected. In other words, learning works because "history repeats itself."

The *mobility model* of a user is a stationary stochastic process $V = V_i$, such that V_i assumes the value $\upsilon_i \in \vartheta$ in the event that the ith update reports the user in zone υ_i. The joint distribution of any subsequence of V_is is invariant with respect to shifts in the time axis, that is,

$$\Pr\left[V_1 = \upsilon_1, V_2 = \upsilon_2, \ldots, V_n = \upsilon_n\right] = \Pr\left[V_{1+l} = \upsilon_1, V_{2+l} = \upsilon_2, \ldots, V_{n+1} = \upsilon_n\right]$$

Zone	Activity
m	Wake up in master bedroom
a	Go to attached bathroom
m	Back in bedroom
c	Change in closet
m	Back in bedroom
r	Out of bedroom
k	Go to kitchen to make breakfast
d	Go to dining room to eat
k	Back to the sink at kitchen
d	Walk to the garage through dining room
g	Start the car at garage
o	Drive away through outdoor area
o	Drive back through outdoor area
g	Back to garage
d	Enter through dining room
k	To kitchen for a snack and drink
l	To living room for watching TV
r	On the way to bedroom
m	In bedroom
c	To closet for changing
m	Back in bedroom
a	To attached bathroom
m	Back in bedroom
r	On the way to more TV watching
l	Back in living room
k	In kitchen to cook dinner
d	Eat dinner
l	Back for the *Tonight Show*
r	On the way to bedroom
m	In bed for the night
a	Wake up and go to bathroom
m	Back in bedroom
c	Change in closet
:	8

Table 1. *Bob's movement history for a day.*

The objective of our update scheme is to learn user mobility, endowing the paging mechanism with a predictive power which reduces average paging cost. Unlike earlier schemes, this provides an on-line algorithm for optimizing the paging problem and allows both the paging process and the update mechanism to be considered.

for every shift l and for all $v_i \in \vartheta$. The movement history is a trajectory or sample path of V.

The goal of our algorithm is to construct a universal predictor or estimator for the user mobility model. Our proposed scheme creates a dictionary of zone ids treated as character symbols and uses the dictionary to gather statistics based on movement history contexts, or phrases. Being motivated by the dictionary-based LZ78 compression algorithm [5], our algorithm [6] assumes the name "LeZi-update" (pronounced "lazy update"), as will be clear in the following.

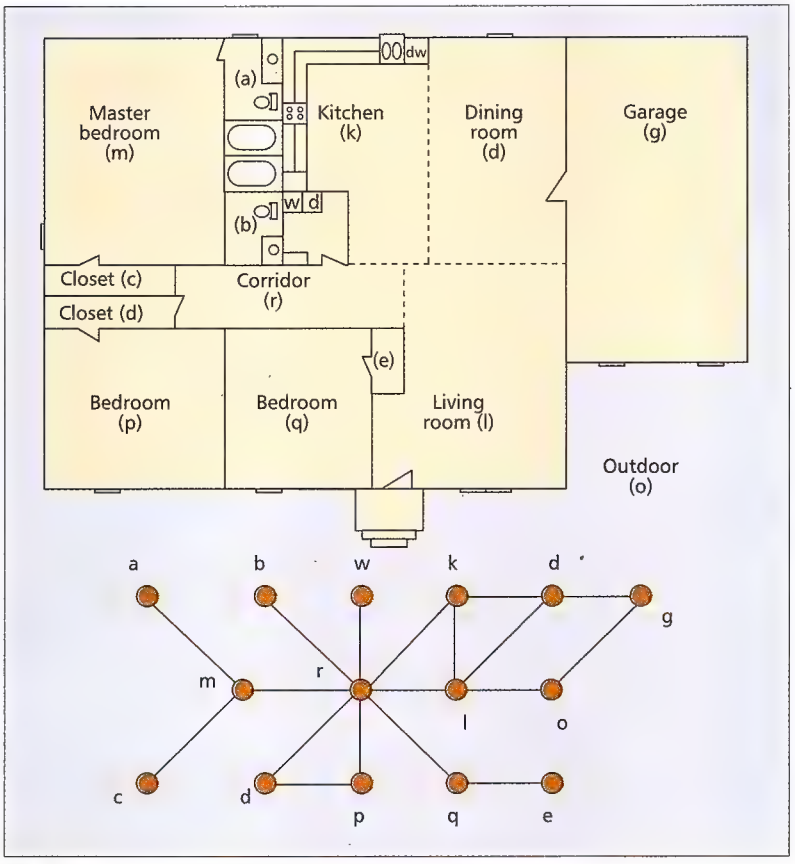


Figure 2. *A graph model of a smart home floor plan.*

LeZi-Update Algorithm

Procedure Encoder

```
loop
  wait for next symbol v
  if(w.v in dictionary)
    w := w.v
  else
    encode <index(w),v>
    add w.v. to dictionary
    w := null
forever
```

Procedure Decoder

```
loop
  wait for next codeword <i,s>
  decode phrase := dictionary [i].s
  add phrase to dictionary
  increment frequency for every
          prefix of phrase
forever
```

Enhanced Decoder

```
loop
  wait for next codeword <i,s>
  decode phrase : = dictionary [i].s
  add phrase to dictionary
  increment frequency for every prefix of every suffix of phrase
forever
```

The LeZi-Update Algorithm

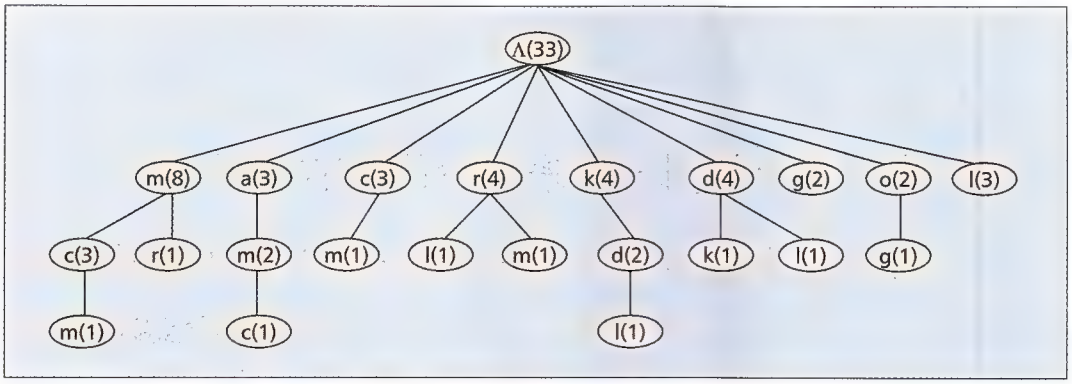
The LeZi-update algorithm (see box this page) enhances an underlying update scheme. The algorithm captures the sampled message, or movement history, and processes it in chunks. Updates are encoded and reported periodically as a sequence "$C(w_1)\ C(w_2)\ C(w_3)\ \ldots$," where the w_is are segments of the movement history and $C(w)$ is the encoding for segment w. The prime requirement for LeZi-update (following LZ78) is that the w_is be distinct.

The system's knowledge about the mobile's location always lags by at most the gap between two updates. The uncertainty increases with this gap, but larger gaps reduce the number of updates. A natural action would be to delay the update if the current string segment being parsed has been seen earlier (i.e., the path traversed since the last update is a familiar one). Although the gap increases, it is expected that the information lag will not affect paging much if the system can make use of the profile generated so far. This prefix-matching technique of parsing is the basis of the LZ78 compression algorithm, which encodes variable-length string segments using fixed-length dictionary indices, while the dictionary gets continuously updated as new phrases are seen. We outline this greedy parsing technique as used in our context. The mobile acts as the *encoder*, while the system takes the role of the *decoder*. For Bob's movement history, the Lempel-Ziv parsing boils down to the phrases *m, a, mc, mr, k, d, kd, g, o, og, dk, l, r, mcm, am, rl, kdl, rm, amc, ...*, where the commas represent the points of updates.

The LZ78 algorithm emerged out of a need to find a universal variable-to-fixed length coding scheme, where the coding process is interlaced with the learning process. The key to the learning is a decorrelating process, which works by efficiently creating and looking up an explicit dictionary. The algorithm in [5] parses the input string $v_1, v_2, \ldots, v_n$, into $s(n)$ distinct substrings $w_1, w_2, \ldots, w_{s(n)}$ such that for all $j \geq 1$, the prefix of substring w_j (i.e., all but the last character of w_j) is equal to some w_i, for $1 \leq i < j$. Because of this prefix property, substrings parsed so far can be efficiently maintained in a multiway tree or *trie*.

We can further improve the performance of the algorithm by augmenting the dictionary with the suffixes of decoded phrases on the decoder (see the box on this page). The enhanced trie for our example takes the shape shown in Fig. 3, where the frequency counts of the phrases are shown within parentheses. The symbol Λ denotes an empty string.

As the process of incremental parsing progresses, increasingly larger phrases accumulate in the dictionary. Consequently, estimates of conditional probabilities for larger contexts start accruing. Intuitively, the process would gather the predictability or richness of higher and higher order Markov models. Since there is a limit to



■ **Figure 3.** *The trie for the enhanced LZ symbol-wise model.*

the model richness for stationary processes, the symbol-wise model should eventually converge to the universal model.

Each user's location database holds a trie, which is the symbol-wise context model corresponding to the enhanced Lempel-Ziv incremental parse tree. Each node except for the root preserves the relevant statistics that can be used to compute total probabilities of contexts as well as conditional probabilities of the symbols based on a given context. A path from the root to any node *w* in the trie represents a context, and the subtrie rooted at *w* reveals the conditional probability model given that context. The paths from the root to the leaves in the Lempel-Ziv trie represent the largest contexts that are not contained in any other contexts.

Next, we assign probabilities for the occurrence of the symbols (zones) on the path segment to be reported by the next update. These segments are the sequences of zones generated when traversing from the root to the leaves of the subtrie representing the current context. The estimated conditional probabilities for all the zones at the current context constitutes the conditional probability distribution.

We use the enhanced trie in Fig. 3 to illustrate a blending strategy known as *exclusion*. Assume that Bob's predicted location is needed; no LeZi-style path update message has been received since receiving the last *amc* in the sequence. The contexts that can be used are suffixes of *amc*, with the exception of itself. First, we need to find all possible paths that can be predicted at these contexts. A list of all such paths are shown in Table 2, with their respective frequencies. (The unconditional probabilities of occurrence of these phrases are then computed by blending. The phrase *m*, for example, appears in the contexts of all the orders 0, 1, and 2. We start from the highest order, i.e., the context *mc*). The phrase *m* occurs only once out of three possible occurrences of this context, the other two producing null prediction. Thus, we can predict the phrase *m* with probability 1/3 at context *mc* and fall back to the lower order with probability 2/3. Now *m* occurs once at context *c*, out of a total of three occurrences of the context. Thus, *m* can be predicted with probability 1/3 at the order-1 context. Due to two occurrences out of three producing null predictions, we need to escape to lower order with probability 2/3. Finally, *m* shows up four times out of 33 possible phrases, leading to a probability value 4/33. The blended probability of phrase *m* is thus 1/3 + 2/3 {1/3 + 2/3 (4/33)} = 0.6094. Since the phrase is made of only one symbol, *m*, the whole probability is assigned to that symbol.

We have shown here that location learning is one prediction task essential for intelligent environments. The LeZi-update algorithm builds user profiles to perform this prediction based on the LZ78 compression algorithm. The effectiveness of this algorithm is apparent for wireless service providers [6], and we are now in the process of demonstrating its use for intelligent environments such as MavHome [9].

We have shown here that location learning is one prediction task essential for intelligent environments. The LeZi-update algorithm builds user profiles to perform this prediction based on the LZ78 compression algorithm.

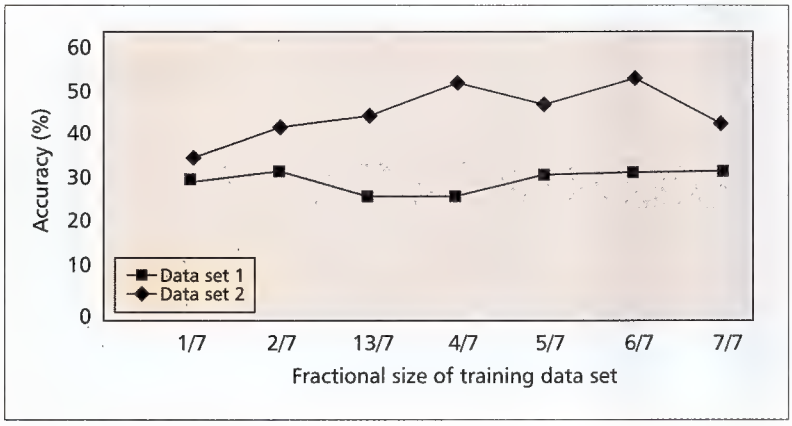
Inhabitant Action Prediction

In the MavHome environment, the intelligent agent representing the house needs to predict the inhabitant's next action in order to automate the routine and repetitive tasks for the inhabitant. Patterns observed in past inhabitant activities can be used to aid the agent decisions for controlling devices throughout the home. In this section we describe this second role of prediction in the MavHome architecture and present a sequence matching approach to performing inhabitant action prediction based on collected histories of actions.

Prediction is a heavily researched area in artificial intelligence. The ONISI system [7] and the UNIX command prediction algorithm [8] employ pattern matching. IDHYS [9] represents an

mc (order-2)	*m* (order-1)	Λ (order-0)		
m|*mc*(1)	*m*|*c*(2)	*m*(4)	*c*(2)	*kdl*(1)
Λ|*mc*(2)	Λ|*c*(2)	*mc*(2)	*cm*(2)	*d*(2)
		mcm(1)	*r*(2)	*dk*(1)
		mr(1)	*rl*(1)	*dl*(1)
		a(1)	*rm*(3)	*g*(2)
		am(1)	*k*(2)	*o*(1)
		amc(1)	*kd*(1)	*og*(1)
				l(3)

■ **Table 2.** *Phrases and their frequencies at contexts* mc, m, *and* Λ.



■ Figure 4. *SHIP predictive accuracy results.*

approach to action prediction based on the Candidate Elimination algorithm.

Our proposed Smart Home Inhabitant Prediction (SHIP) algorithm matches the most recent sequence of events with sequences in collected histories. SHIP considers matches of length three or greater, and returns matches of sufficiently high value based on the *match frequency* (number of occurrences of the pattern in the history) and *match length* (length of the matched sequence).

In the SHIP algorithm, the inhabitant commands are encapsulated using *actions* and *matches*. When the inhabitant issues a command to a device, it is recorded as an action in the inhabitant *history*. A match identifies a sequence in history that matches the immediate event history (a sequence ending with the most recent event). The SHIP's predicted action corresponds to the action that followed the matched sequence most frequently in the inhabitant history. A match queue is maintained to ensure a match time that is close to linear.

The SHIP algorithm consists of two steps. First, the match queue is updated when a new action is recorded. At time t in state s we compute $l_t(s, a)$, the length of the longest sequences that end with action a in state s and match the history sequence immediately prior to time t. In addition, we define a frequency measure $f(s, a)$ that represents the number of times the action a has been taken from the current state. In the second step, the matches in the queue are evaluated based on match length and frequency.

To allow for gradual changes in inhabitant patterns over time, the value of a matched pattern can be multiplied by a user-specified decay factor. The user also has the flexibility to weight the match length and match frequency factors that affect a match value. Because an inhabitant pattern is likely to contain small variations between occurrences, an inexact match is employed to find sequence matches. The user can specify an *inexact threshold* which represents the maximum percentage of mismatches that may occur in a matched sequence.

To test the predictive accuracy of the SHIP algorithm, we equipped devices in several homes with X10 controllers and collected action histories for these households with different numbers of people, types of activities, and spans of time. For more precise experiments we created a synthetic data generator that simulates several possible activity scenarios. With each type of activity is associated a Gaussian probability distribution over start times and durations from which actual event histories can be generated.

The graph in Fig. 4 shows the percentage of correct predictions for two of the data sets. For this experiment a decay factor of 0.0 is used, and the match frequency and length weights are set to 1.0. These data sets reflect activities collected over a maximum period of 30 days. Data set 1 captures 747 activities of four inhabitants with 11 devices, and data set 2 represents 3000 activities of one inhabitant using 16 devices.

As the figure shows, the accuracy of the algorithm generally improves as the amount of training data, calculated as a percentage of the total history, increases. The greatest predictive accuracy for data set 1 is 33.3 percent, and for data set 2 53.4 percent. The accuracy rate reflects greatly improved performance over a random guess. As a separate test, we measure the predictive accuracy made by one of the top three matches identified by SHIP. In this case, the predictive accuracy is over 80 percent. These results indicate that the SHIP algorithm is effective for predicting inhabitant actions in a smart home, particularly if the top several matches are considered.

Discovery of Significant Patterns

The SHIP algorithm is useful in identifying likely activities of a smart home inhabitant. This information can be used to automate interactions with the home, removing the need for manual control of devices. A wrong prediction, however, can be annoying or detrimental if the inhabitant must undo the action executed by the house or repair damage caused by a faulty decision.

Instead of identifying and automating each inhabitant pattern, we describe here a prediction algorithm, called Episode Discovery (ED), that identifies *significant episodes* within an inhabitant event history. A significant episode can be viewed as a related set of device events that may be ordered, partially ordered, or unordered. A significant episode occurs at some regular interval or in response to other significant episodes called *triggers*. The goal of the intelligence framework in our problem domain is to mine the input stream in order to discover the significant episodes. Actions can then be automated based on the significance of the discovered pattern as well as the predictive accuracy of the next event.

Our approach is based on the work due to [11] for mining sequential patterns from time-ordered transactions. Our home automation problem differs from previous research in that the input sequence does not consist of explicit transactions, but merely interactions with home devices. Unlike the previous sequence mining problem, the significant episodes in an intelligent environment may be ordered (sequential) or unordered (member of a set). In addition, many of the episodes in our environment will occur on a regular basis (daily, weekly) and need to be recognized for this regularity. In our

MavHome scenario, the following device activity sequences occur on a regular basis and should be detected by our algorithm:

- HeatOn (daily)
- AlarmOn, AlarmOff, BedroomLightOn, CoffeeMakerOn, BathRoomLightOn, BathRoomVideoOn, ShowerOn, HeatOff (daily)
- BedroomLightOff, BathRoomLightOff, BathRoomVideoOff, ShowerOff, KitchenLightOn, KitchenScreenOn (daily)
- CoffeeMakerOff, KitchenLightOff, KitchenScreenOff (daily)
- HotTubOn (daily)
- HotTubOff (daily)
- SprinklerOn (weekly)
- SprinklerOff (weekly)
- VCROn (weekly)
- VCROff (weekly)
- OrderGroceries (weekly)

Other activities, such as robot activation, would not be identified by ED as significant because they do not occur with any predictable regularity.

To mine the data, the input sequence is partitioned into transaction-like collections of events by sliding a window over the event history and viewing the collection of events within the window as an unordered set. The minimum description length (MDL) principle is used to evaluate potential sequences. The MDL principle targets patterns that can be used to minimize the description length of the database by replacing each instance of the pattern with a pointer to the pattern definition. This evaluation measure thus identifies patterns that balance frequency with pattern length. As a result, automating these sequences will significantly reduce the amount of necessary interaction between an inhabitant and the environment. Another feature of ED is that patterns are evaluated for day, week, and month regularity as well as MDL value. Significant episodes will then be selected based on the overall evaluation measure, and used as the basis for activity prediction and home automation.

The episode discovery problem is defined as follows. Let E be the set of all device events. An *event occurrence* O is a pair (e, t) relating an event e to an integer time value t. An *event sequence* S is an ordered sequence of event occurrences. We define an *episode* ε as a set of event occurrences, and a *candidate itemset* I as a set of events and episodes, $I = (\{e_1, e_2, e_3, e_n\}, \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\})$, where each event e_i has an occurrence in each of the episodes ε_j, for $1 \le i \le n$ and $1 \le j \le m$. A significant episode L is an episode that meets or exceeds the evaluation threshold, and an *event sequence description* D is a description of an event sequence using significant episodes and event occurrences. ED operates as follows:

1 Construct an event sequence S from input O.
2 Partition S into episodes using a sliding window of length w.
3 Create candidate itemsets from the episodes.
4 Compute compression values for each of the candidate itemsets.
5 Using a greedy approach, identify the candidate itemset that minimizes the description length of the set of episodes as a significant episode.
6 Remove all of the episodes associated with the candidate itemset from the remaining candidate itemsets.
7 Remove all candidate itemsets that have an empty episode set.
8 Repeat steps 4–7 until the list of candidate episodes is empty.

An online incremental version of ED has been implemented using C++. To validate this algorithm, a synthetic 28-day data set was generated reflecting our MavHome scenario. To make the problem more challenging, noisy events were included, episodes were overlapped, and the times of the event occurrences were varied each day. Using a 15 minute time window, ED discovers the 11 significant episodes described by the scenario. Using a 12 hour window, the daily and weekly occurrences are detected as separate significant episodes. These results show that ED can be used to aid in the automation of device interactions, as described by our MavHome scenario.

Using a 12 hour window, the daily and weekly occurrences are detected as separate significant episodes. These results show that ED can be used to aid in the automation of device interactions, as described by our MavHome scenario.

Conclusions

In this article we present the MavHome smart home architecture, which allows a smart home (or other intelligent environment) to act as a rational agent. As a rational agent the home receives input from sensors and selects an appropriate action that is executed through the use of effectors. This architecture allows the integration of research in machine learning, databases, mobile computing, robotics, and multimedia computing that is essential for smart home development.

As part of the MavHome architecture, several prediction algorithms are introduced that play critical roles in an adaptive and automated environment such as MavHome. The first prediction algorithm, LeZi-update, provides an optimal approach to the location management problem that is useful in determining the position of an inhabitant for routing messages and multimedia information. This novel approach uses movement histories to learn likely future locations. The second algorithm, SHIP, uses sequence matching with inexact allowances and decay factors to determine the most likely next inhabitant interaction with the home. Results from synthetic and real collected smart home data indicate that the predictive accuracy is high even in the presence of many possible activities. The final algorithm, ED, uses the principle of minimum description length to determine which episodes in an inhabitant history are significant. Significance is determined based on the ability to compress the description length of the history as well as periodicity. As a result, these episodes represent events that should be automated by MavHome.

We have demonstrated the effectiveness of these algorithms on collected data. The next step for this effort will be to implement the architecture in the context of actual smart environments. In these contexts we will show the effectiveness of the MavHome architecture operating as a rational agent, and its ability to improve the lifestyle of inhabitants in a variety of intelligent environments.

We have demonstrated the effectiveness of these algorithms on collected data. The next step for this effort will be to implement the architecture in the context of actual smart environments.

References

[1] I. F. Akyildiz and J. S. M. Ho, "Dynamic Mobile User Location Update for Wireless PCS Networks," *WL Nets.*, vol. 1, no. 2, 1995, pp. 187–96.

[2] I. F. Akyildiz and J. S. M. Ho, "Movement-Based Location Update and Selective Paging for PCS Networks," *IEEE Trans. Net.*, vol. 4, no. 4, 1995, pp. 629–38.

[3] Y. Birk and Y. Nachman, "Using Direction and Elapsed-Time Information to Reduce the Wireless Cost of Location Mobile Units in Cellular Networks," *WL Nets.*, vol. 1, no. 4, 1995, pp. 403–12.

[4] G. P. Pollini and C.-L., "A Profile-Based Location Strategy and Its Performance," *IEEE JSAC*, vol. 15, no. 8, 1997, pp. 229–32.

[5] J. Ziv and A. Lempel, "Compression of Individual Sequences via Variable-Rate Coding," *IEEE Trans. Info. Theory*, vol. 24, no. 5, 1978, pp. 530–36.

[6] A Bhattacharya and S. K. Das, "LeZi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks," *Proc. ACM/IEEE Int'l. Conf. Mobile Comp. and Net.*, 1999, pp. 1–12.

[7] P. Gorniak and D. Poole, "Predicting Future User Actions by Observing Unmodified Applications," *Nat'l. Conf. AI*, 2000.

[8] B. Korvemaker and R. Greiner, "Predicting UNIX Command Lines: Adjusting to User Patterns," *Proc. Conf. Intelligent Apps. AI*, 2000.

[9] A. Roy *et al.*, "Location Aware Resource Management in Smart Homes," submitted for publication, Nov. 2002.

[10] J.-D. Ruvini and C. Dony, "APE : Learning Users Habits to Automate Repetitive Tasks," *Int'l. ACM Conf. Intelligent User Interfaces*, 2000, pp. 229–32.

[11] R. Srikant and R. Agrawal, "Mining Sequential Patterns," *Proc. 5th Int'l. Conf. Extending Database Tech.*, 1996.

Biographies

SAJAL K. DAS (das@cse.uta.edu) received a B.Tech. in 1983 from Calcutta University, an M.S. in 1984 from the Indian Institute of Science, and a Ph.D. in 1988 from the University of Central Florida. Currently he is a professor of computer science and engineering and director of the CReWMaN Center at the University of Texas at Arlington. His research interests include wireless networks, mobile and pervasive computing, parallel/distributed processing, performance modeling, and simulation. He has published over 200 papers in these areas and holds four U.S. patents.

DIANE COOK (cook@cse.uta.edu) is currently a professor in the Computer Science and Engineering Department at the University of Texas at Arlington. Her research interests include artificial intelligence, machine learning, data mining, robotics, and parallel algorithms for artificial intelligence. She has published over 120 papers in these areas. She received her B.S. from Wheaton College in 1985, and her M.S. and Ph.D. from the University of Illinois in 1987 and 1990, respectively.

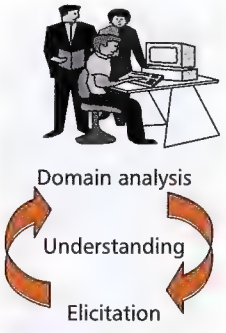
AMIYA BHATTACHARYA [StM] (bhatt@cse.uta.edu) is currently a Ph.D. candidate in the Department of Computer Science and Engineering at the University of Texas at Arlington. He received his B.Tech. and M.Tech. degrees in 1987 and 1989, respectively, from the Indian Institute of Technology, and his M.S. in computer science in 1991 from the University of California, San Diego. His research interests include mobile computing and communication systems, network protocols, measures for performance and dependability, optimization in dynamic systems, and the application of randomized algorithms, online algorithms and information theory. He is a student member of ACM and ACM SIGMOBILE.

EDWIN O. HEIERMAN III (heierman@cse.uta.edu) is a member of the research development staff at Abbott Laboratories Diagnostic Division, Irving, Texas. His interests are in the areas of embedded software development, Internet appliances, and knowledge discovery in databases. He received a B.S. degree from the United States Air Force Academy in 1984 and an M.C.S degree from the University of Texas at Arlington in 1997, and is currently pursing a Ph.D. at the University of Texas at Arlington.

TZE-YUN LIN (tylin@cse.uta.edu) came to the United States as an exchange student. She received her B.S. in computer science in 1999 from Texas Christian University, Fort Worth, Texas, and is currently an M.S. student in the Department of Computer Science at the University of Texas at Arlington. Her research interests include artificial intelligence, machine learning, and utility reasoning.

A HYBRID ANALYSIS AND ARCHITECTURAL DESIGN METHOD FOR DEVELOPMENT OF SMART HOME COMPONENTS

JOHN R. DURRETT, TEXAS TECH UNIVERSITY
LISA J. BURNELL, TEXAS CHRISTIAN UNIVERSITY
JOHN W. PRIEST, UNIVERSITY OF TEXAS AT ARLINGTON



Accurate requirements provide the foundation for successful product development. Determining accurate requirements is difficult in many cases and is especially problematic in new and rapidly evolving domains. Development of Smart Home technologies provides such a challenge.

ABSTRACT

Accurate requirements provide the foundation for successful product development. Determining accurate requirements is difficult in many cases and is especially problematic in new and rapidly evolving domains. Development of smart home technologies provides such a challenge due to their dynamic design environment consisting of emergent technology, with minimal existing systems to evaluate, few standards, and users with vague ideas of the benefits or the possibilities. The most troubling problem, however, is how to optimize user satisfaction considering the wide range of user types and preferences and the dynamic system environment created by constant introduction of new products. For early development in this user-focused environment, we propose a new approach based on the synthesis of well established techniques from software engineering, management theory, and hardware product development. We propose the fusion of the use case and house of quality analysis models, simulators, and prototypes, and information processing theory coordination techniques.

INTRODUCTION

Effective product development is a requirements-driven process. The keys to successful product development are to know the customer's current needs and to provide a product that meets these needs and is easily adaptable to future requirements at a competitive cost and in a timely manner. Customer requirements should include items for which the customer recognizes a need as well as items for which the customer does not because of a lack of information concerning new technologies or innovative ideas. These customer requirements, along with technologies, prices, and available alternatives, are always evolving.

When analysts are determining system requirements, a primary concern is communication between domain expert and system designer. A primary focus of our ongoing research efforts is the synthesis of well established techniques from software engineering, management theory, and hardware product development in an effort to facilitate communication in the requirements determination process. Our intention is to show that a synthesis of:

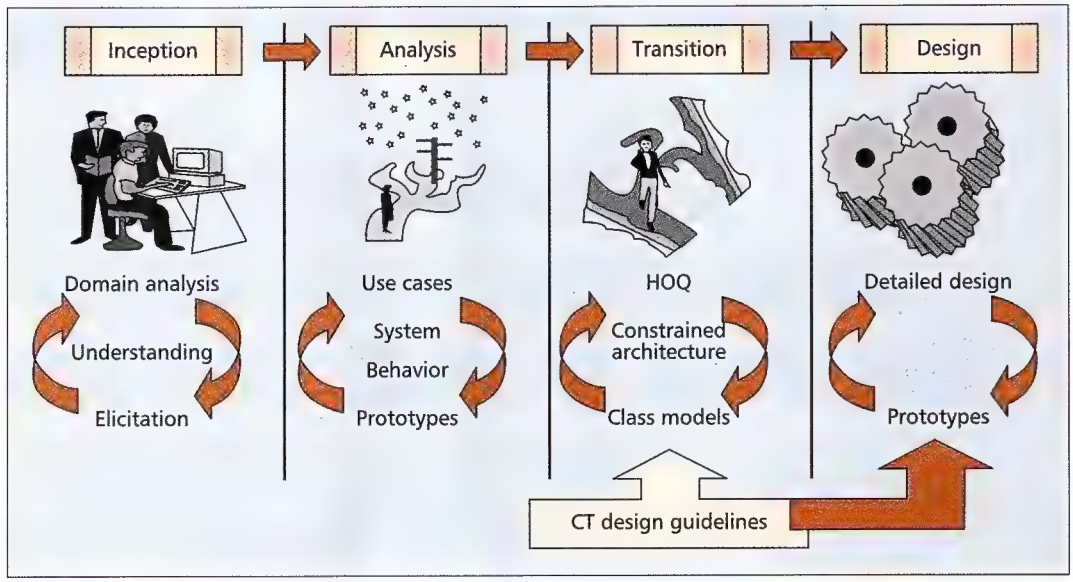
- Human employee coordination techniques from information processing theory (IPT)
- Object-oriented analysis and design techniques such as use cases and prototyping
- Hardware product development methods that use house of quality (HOQ) will increase the effectiveness of any systems design project, especially one in a dynamic incipient field such as the smart home.

The anticipated result is the creation of smart home components that can effectively and efficiently satisfy customers' needs as part of an "integrated" smart home environment.

Requirements definition is the process of identifying, evaluating, quantifying, prioritizing, and documenting specific needs for the development of product requirements for a new product, process, or service [1]. This process is typically very costly and critical to overall project success. Since consumers are not familiar with smart home technologies or their capabilities, one of the fundamental issues facing smart home developers is determining what potential users really want and need. Many imaginative technologies are ready, or nearly ready, to be inserted into the average middle-class home. The danger is that designers will use their own expertise in creating products for much less technically savvy consumers. It is very important to the success of smart home products that the solution match the requirements and that designers not fall into the same trap that flawed early information systems. For example, does a consumer really want a thermostat with 56 different defaults [2]? What product characteristics are needed to convince the consumer and home builder to purchase these products? Moreover, the products developed must work with existing and future products to form an integrated home-wide architecture.

1536-1284/02/$17.00

It is our premise that organizational control and coordination structures are also effective metaphors for designing system architectures for distributed multi-agent environments, such as the Smart Home. Applying these concepts yields distributed, task-oriented architectures that are robust and adaptable.



■ **Figure 1.** *Smart home early development process.*

Even product designers cannot agree on what capabilities the future Smart Home will have. A 1999 Defense Advanced Research Projects Agency (DARPA) Workshop on Smart Spaces identified one research issue as "what we can and should expect for human interaction with our environment." The smart home designer needs quantified design requirements on customer issues such as value, usability, customization, privacy, security, cost, maintenance (corrective and predictive), and adaptability. Use cases are used to promote communication and understanding between designer and consumer. Prototyping and simulation are used to visualize system behavior, confirm resulting designs, and provide initial documentation to designers. HOQ is used to further refine system attributes and to align customer priorities with proposed design models. Contingency theory (CT) and IPT are used to gain an initial understanding of the required system structure given the domain in which it operates. The utilization of these methods in the four phases of the software development process for the smart home is shown in Fig. 1.

In the inception phase [3], the primary activities are to understand the domain and customer and to perform domain analysis and requirements elicitation. During the analysis phase, the goal is to describe the system behavior: what the proposed system is to do. In object-oriented development, the line between analysis and design is not as well defined as in structured methods. Use case modeling is used to capture system behavior; class modeling defines classes, attributes, and their relationships; and dynamic modeling determines the interactions between classes. The transition from analysis into design, or from external behavior to internal structure, proceeds with CT/IPT informing the structuring of the design. These ideas are captured within the HOQ and class models (interaction diagrams may be used here for complex internal behavior). This process leads to detailed design, where again prototypes are used, although for different purposes than in analysis. In object-oriented design, the class methods are defined.

Contingency Theory/ Information Processing Theory

Organizational theory (OT) is a field of study that examines an organization's structure, constituencies, processes, and operational results in an effort to understand the relationships involved in creating effective and efficient systems [4]. A postulate of OT, CT, is that although there are many ways to organize due to the unpredictability and diversity of the task environment, not all are equally effective. An extension to CT, IPT, postulates that this environment-determines-organizational-structure dependency is the result of the coordination requirements among the basic human constituent elements and tasks that make up the organization [5]. This relationship exists because a primary method for adapting the organization to its environment is through grouping, coordinating, and controlling human employees. In simple, slowly changing environments, highly structured organizations exist and are successful. This organization structure works because over time rules develop that can be encoded into organizational policy. The need for rapid adaptation to environmental changes is rare. However, organizations that compete in rapidly changing, complex environments do not have the luxury of standard operating policies. Organizational communication and control must be dynamic and flexible. For these organizations to survive, employees must be empowered to make decisions and adapt to environmental demands. It is our premise that organizational control and coordination structures are also effective metaphors for designing system architectures for distributed multi-agent environments, such as the smart home. Applying these concepts yields distributed, task-oriented architectures that are robust and adaptable.

Our Example System

To illustrate this methodology a smart entertainment controller (SEC) is used as an example. The SEC is intended to be a highly intelligent, easy-to-use system that controls typical entertainment

room devices such as a TV, VCR, DVD, CD player, PC, telephone, and lighting systems. The system accommodates multiple users, learns inhabitant preferences, and coordinates with other smart home devices. An example scenario follows:

A person who lives in the home walks into the room at the time the person usually gets home from work. The television automatically turns on and tunes to a channel the person often watches at this time every workday. After determining that the show is a repeat and that the person has already seen it, the system notifies the person. The person decides not to see this episode again and tells the system. The system changes the television to another station showing a program the system has learned the person will like.

The SEC's main function is to identify an inhabitant and control the entertainment devices based on the inhabitant. The SEC maintains a database to track each inhabitant's activities and learn corresponding habits and preferences in order to automate actions based on expected and observed behavior. Like many smart home appliances, the SEC software must be customizable for the diverse population but simple enough for a layperson to use it. Artificial intelligence (AI) techniques such as case-based reasoning and reinforcement learning could ease the user's burden in configuring and modifying the behavior of the appliance. But where and at what level should these techniques be applied (i.e., where will they create the most value)? How can the architecture support new learning and inhabitant identification capabilities without inconveniencing the inhabitants?

We now discuss how we combine the underlying methods in our hybrid approach to develop requirements and the architectural design. This discussion includes an overview of each of the methods and example artifacts developed for the SEC.

Inception: Understanding the Domain (Domain Analysis) and Soliciting Requirements

An analyst's first step is to understand the domain by identifying customer wants and needs. Several approaches are feasible during this process. The design team should use several of these methods to ensure that the final requirements are representative of the customer. Methods and sources for capturing and documenting customer needs include but are not limited to [1]:

- Customer interviews, including techniques such as surveys and focus groups
- Design partnerships where the company's customers participate in the design process
- Data mining of customer sales, preference, and Internet information
- Consultants or experts who specialize in identifying what customers want
- Brainstorming sessions that help generate innovative ideas that change the norm for this type of product
- Personal and company experience including previous successes and failures
- Published information such as magazines, Internet sites, and patents
- Market and competitor benchmark analysis identifying best in class and innovative ideas
- Prototyping and virtual reality simulators to study customer responses
- Systematic methods such as house of quality or quality function deployment

MavHome	AMX [6]	SEI/CMU	Honeywell [7]
Comfort*	Powerful	Usability	Communications
Productivity*	Versatile	Performance	Security
Cost	Flexible/adaptable*	Modifiability*	Programming
Security	Fast	Security	Adaptability*
	Easy to install	Safety	Intelligence
	Easy to program	Reliability	
	Easy to use	Interoperability*	
	Easy to upgrade*		
	Easy to troubleshoot		

Table 1. *Smart home technology feature lists. *: Adaptability features.*

Defining customer needs is a complex process resulting in many different and conflicting types of information. This complexity is especially true for breakthrough products such as the smart home. The customer may not understand the implications of new technology breakthroughs or the need to integrate disparate technologies. One begins by capturing some abstract, vaguely defined features. Examples are shown in Table 1.

Although these high-level feature lists are an important starting point, they do not provide the quantified details necessary to ensure success. Each feature must be refined prior to use case capture to guide that process. To better illustrate our approach, we will focus on the adaptability features, marked by asterisks in Table 1. Adaptability can be defined as adjusting to a new situation. A second level of refinement for the adaptability feature is shown in Table 2. As part of use case capture, prototypes and simulators are used to visualize and further refine the behavior manifested by these features.

Analysis: Analyzing and Specifying Requirements

From the domain understanding and initial feature analysis, the translation into requirements specification begins. This requirements process begins with use case modeling.

Use Cases

Use cases provide a framework or viewpoint of potential capabilities to allow users to visualize what a new product might do for them. The goal is to convey what the system "can" do, not "how" it will be done. Use cases work well in the design situation because they focus on how the user is to interact with the system (Fig. 2) and allow validation by independent experts. A use case model is a semi-formal technique that provides a bridge between the customer's need for informality and the developer's need for precision.

Use cases explain how a product will be used for a particular application or task. A product will have several use cases depending on the number of product uses, features, and different users. Use cases are developed to illustrate the desired functionality and problem-solving focus for each feature. The initial use cases help the

High-level feature	Catalyst for adaptability	Refinement
Adaptability	Devices	Easy to install and control "new" devices Easy to install and control devices using "new" technologies
	User capability (e.g., "new"users)	Interfaces support both novice and experienced users
	User preferences	Easy to change user preferences Automates functions based on user preference predictions
	Data availability	Uses available data for interfaces and preferences (e.g., "new" TV schedules)

Table 2. *Second-level refinements for the adaptability feature.*

development team learn system requirements and provide the basis for early prototyping. Scenarios, which specify detailed behavior within a use case, are later decomposed to define task responsibilities and constraints, leading to class and interaction models. As scenarios become more detailed, they are used for developing design and test specifications, and identifying typical and atypical process flows within the system. These analyses are elicited using a number of techniques, including interactive observations, structured interviews, demonstrations, and focus groups. Scenario descriptions include responsibilities, capabilities, and views. The scenario perspective focuses on each user's viewpoint of what activity is taking place, rather than how or what technology is used to support it.

Use case models, used alone, are problematic for the smart home domain. Use case scenarios can be tedious to write out. A quick sketch, as in an XP user story, with a link to a prototype to show behavior of the system, is more acceptable and expressive, especially for highly interactive or real-time systems.

USE CASE: SEC Automated Preference Selection

Feature Category: Adaptability
Description: Automate functions based on user preference prediction.
Actors: Inhabitant

Primary Scenario:
1. System activates at time intervals and at inhabitant recognition.
2. System retrieves preference model, if available, for the inhabitant appropriate for the current time, date, and day of week.
3. System compares preference model to current device settings.
4. System offers inhabitant opportunity to change current settings to those specified in preference model.
5. Inhabitant accepts proposed changes.
6. System alters current device settings to those in the preference model.

Alternative scenario 1: No preference model exists for inhabitant.
1. Steps 3 through 6 are skipped.

Alternative scenario 2: User declines proposed changes at step 5.
1. Steps 5 and 6 are skipped.

Alternative scenario 3: Current state matches those in preference model.
1. Steps 4–6 are skipped.

Notes: Default time interval is 30 minutes (on the half-hour). Inhabitant may change this time interval or deactivate this function.

Figure 2. *A simplified use case example (single inhabitant behavior is shown for simplicity).*

Prototypes and Simulators

The basic philosophy behind the use of prototyping is to provide a communications platform for studying consumer needs and refining product Use Case scenarios. This focus can be especially helpful for Smart Home technologies since the product is very innovative. Prototyping helps the design team by encouraging the consideration of as many product development issues as possible during the early phases and by identifying potential problems. The steps are [1]:

- Produce prototypes that provide the reviewer a realistic view of the proposed design
- Continuously produce prototypes throughout the product development process
- Show the prototypes to everyone

Simulators are used for developing and verifying requirements models. Customers cannot imagine what they need when the possibilities are so radically different from their normal experience. The simulator immerses them in a virtual world so that they can experience how life could be. From this experience, a customer can then report likes and dislikes that can be transformed into a prioritized set of requirements. For example, using traditional online TV guides to mark a show to record is a much different experience from using a system that automatically records a show about the making of Star Trek because it "knows" you are a big fan. Furthermore, if the system notifies you of its actions on Friday night because that is when you like to relax with a TV show and pizza, the experience has moved even further from the norm.

The SEC simulator sends commands to the SEC and simulates the output. A graphical environment shows the results of the SEC's commands to the various devices in the room. The simulator allows engineers to create scripts to test SEC functionality. A script is a temporally ordered list of commands that are sent to the SEC in succession. Reports are then generated to show how the SEC reacted to each command in the script and to show various statistics that can be used for learning. The simulator can receive and send messages to virtual or physical devices. The Knowledge Query and Manipulation Language is currently used for communication.

Rapid prototyping as a specification technique offers fast, unambiguous specification of requirements (embodied in the prototype). Schach [8] reports on experiments of Boehm, Gray, and Seewaldt (1984) in which seven different versions of a product were compared. Four were specified, and three were prototyped. The results showed that prototyping and specifying yielded equivalent performance. But while the prototyped versions required 40 percent less code and 45 percent less effort and were easier to learn and use, they were inferior to specification in terms of functionality, robustness, and integratability. Moreover, specifications are needed for testing, maintenance, and contract purposes. The specification model consists of a use case model, HOQ, and supplemental requirements documents for capturing nonfunctional and other overall system requirements. The prototype, working within the simulator, supports the development and testing of the requirements specification model.

Transition: Mapping Behavior to Design

In transition, we move forward from a detailed specification of system behavior, as captured in the use case model and prototypes, toward object-oriented class models. We facilitate this transition by adapting CT/IPT principles to software development and construction of an HOQ.

House of Quality

After the previous techniques have provided the framework, the next step is to better define customer attributes and priorities, and begin to constrain and quantify design requirements. HOQ or quality function deployment (QFD) is both a requirement definition and conceptual design tool that systematically documents customer needs, benchmarks, competitors, and other aspects, and then transforms this information into design requirements [9]. The steps are:

- Determine and rank customer attributes.
- Document customer perceptions of how well different products meet these attributes.
- Determine "measurable" design characteristics/parameters and rate their relationship to the customer attributes.
- Determine objective requirements and measures (goals) for the design characteristics.

An important result is the identification of critical factors for customer quality. The process starts with the feature table (Table 2) and the use case model. A portion of the resulting HOQ that illustrates adaptability is shown in Fig. 3.

This method requires a large amount of effort to perform correctly. However, Elboushi and Sherif [10] successfully used QFD (documented as an HOQ) followed by prototyping of the user interface to develop the requirements specification for developing a portion of the Advanced Multimission Operations System. They report a 50 percent reduction in requirements capture time, 60 percent improvement in identifying conflicting requirements or potential bottlenecks, and 70 percent increase in consistency. They followed this effort with classical OOD methods, using CRC (Class-Responsibility-Collaborator) cards and additional prototyping to develop the design. Their discussion of how to perform subsystem grouping of classes using "personality" meshes well with our defined design guidelines based on CT.

How CT/IPT Supports the Transition from Analysis to Design

CT/IPT states that the adaptations that organizations, human or software, may utilize to solve the information processing requirements in an organizational structure can be generalized into two broad categories: planning and mutual adjustment. The more heterogeneous, unpredictable, and dependent on other environmental resources a task is, the greater the information processing the organization must be able to do in order to successfully accomplish the task. As diversity and unpredictability increase, uncertainty increases due to incomplete information. As resources, processes, or outputs increase, interprocess coordination requirements and system complexity increase. As uncertainty increases, information processing requirements increase because of management's inability to predict every situation, and more mutual adjustment and cooperation are required. Conversely, with more homogeneous, predictable, and independent tasks, management's ability to predict situations and plan for solutions increases. Thus, standard operating procedures (SOP) can be implemented and hierarchical control systems created to manage the organization effectively. Table 3 shows these environment-architecture relationships.

The SEC design must account for and adapt to the expected variability and complexity in user requirements, rapid technology evolution, and expected demand for mass customization of components. CT is employed to guide the development of the initial architecture, the organization of the components, and the grouping and distribution of tasks to agents.

In previous research we have developed and tested the following CT-based guidelines for creating intelligent agent-based systems:

- Describe business activity and identify tasks: Using the information from the use cases, HOQ, and simulations developers can refine the overall purpose of the software being designed.
- Determine task predictability: Since a basic premise of CT is that the control structure of a business process must match the environment in which it operates, we must identify the predictability of each task.
- Assign tasks to components (objects, agents, etc.): Once the level of predictability has been estimated for each task, the granularity of the components being created can be determined and component designs finalized.
- Group components into teams: As with human organizations, our components can be grouped along any of several dimensions, including task, workflow, product, manager, or communication requirements, as required by the operating environment.
- Identify communications needs: Once teams are determined, the communication requirements of individual components, and of teams, can be determined.
- Construct management groups: In software systems operating in a dynamic environment, management is required only when components are unable to handle events.

Given that our task environment is dynamic (rapidly changing with new equipment introduced, new demands by users, new systems controlled, and components added) and highly complex (lots of different inputs from users and other system components and a wide variety of user and system-level outputs required), IPT guidelines suggest that we have:

- Task-level components, which control very specific activities such as TV viewing or retrieving current program availabilities from a TV Guide.
- "Management-level" components that act at a higher system level. Let us define management in the realm of our SEC as a component, or set of components forming a team, that possesses the following capabilities:

—Monitoring system status and reporting problems, usage, costs, and other statistics

The SEC design must account for and adapt to the expected variability and complexity in user requirements, rapid technology evolution, and expected demand for mass customization of components.

—Learning user preferences
—Handling external resource requirements (e.g., logging on to satellite radio stations or online music databases)
—Handling unexpected problems and emergencies

Management teams are responsible for implementing "plug-and-play," for monitoring the availability of system inputs such as cable or Internet and adapting suggestions based on unexpected outages, for monitoring user logons

			Engineering characteristics				
			Design attributes (measurement units)				
		Importance ranking for adaptabiity (10 most important)	# of users/ preferences (# of users)	Predictive user/ preference accuracy (% accuracy)	Multievent storage (date, time, user, mode) (# of attributes)	# of steps to set/modify preference (# of steps)	"Drop in" user/ language/ culture module (# of modules)
User preferences	Learned preferences, predict what I want	10	X	PP	PP		PP
	Easy to modify existing preferences	9				PP	
User interface capability	Common controls for different devices	7					XX
	Automatically detect and identify user	6	XX	X			
	Novice and experienced users	10		XX	PP	P	PP
	Multiple control interfaces	7	X				
Devices	Wide range of electronics and appliances	7		X	XX	X	XX
	Easy to install new devices	8					XX
Data availability	Many devices changing simultaneously	5	X	X	X	P	
	Access new information (e.g., TV schedules)	8		XX	XX		

Continued from above

		Importance ranking for adaptabiity (10 most important)	# of steps override command (# of steps)	# of user interface modes accuracy (# of modes)	Number of devices controlled (# of devices)	# of protocols/ standards supported (#)	Automatic device installation utility (Yes/No)
User preferences	Learned preferences, Predict what I want	10			X		
	Easy to modify existing preferences	9	P	XX	X		
User interface capability	Common controls for different devices	7			X		PP
	Automatically detect and identify user	6		X			
	Novice and experienced users	10	X	PP	XX		X
	Multiple control interfaces	7		PP	P		X
Devices	Wide range of electronics and appliances	7	X	XX	PP	PP	PP
	Easy to install new devices	8		XX	PP	PP	PP
Data availability	Many devices changing simultaneously	5			XX	X	
	Access new information (e.g., TV schedules)	8					

Relationships: PP Strong positive; P Medium positive; X Medium negative; XX Strong negative

Figure 3. *House of quality for the adaptability feature*

Problem/initial state	Nature of task environment	Model solution
Typical business/application	Stable or slowly changing environment	Hierarchical structure and SOP
Increased number of "exceptional" situations and runtime exceptions	Environment rate of change increasing	"Management" software teams such as learning modules to handle exceptions
Dramatic number of runtime exceptions overloading hierarchy	Dynamic: very high rate of change in task environment	Lateral component coordination through "rules"
Overloaded lateral structure due to high coordination requirements	Dynamic with low to moderate complexity	Self-contained management components to relieve coordination requirements
Overloaded lateral structure due to high coordination requirements	Dynamic with high complexity	Slack resources through use of "cloned" tasks and management components

■ **Table 3.** *Environment-structure relationship.*

and implementing security lockouts for adult-level entertainment, or (if a security module is included in the controller) for reporting unauthorized removal of hardware components to some external authority. As system complexity increases in the form of additional inputs and outputs required of the SEC, IPT suggests that some of the management capabilities should be passed down to the individual task components. For instance, each device manager would contain an individually tailored learning module rather than requiring a centralized module that learns all watched activities.

Conclusions and Future Work

In this and other projects to which we have applied our approach, we have reached the following important conclusions:

- That traditional methods of early design should change to accommodate the dynamic and complex environment of the smart home. This shift will be necessary to allow a product to be both successful as an individual product and to provide the adaptability for the optimization of the entire house.
- That prototyping is essential early in the process. Developers should not expend too much time on use cases/HOQ at first. Instead, they should develop rough models and start the prototype and its simulator. Then use the simulator to refine the requirements models.
- That the organizational metaphor for distributed systems development is useful not only as a design tool, but to communicate with both customers and developers.

Smart home projects need agile software development methods because the domain is complex, rapidly changing, and fast-paced. This need for agility will be especially true in the future with the introduction of smart appliances (refrigerators, remotes, ovens, etc.). Our approach fits within any of the proposed agile software development methods to provide guidance in requirements capture, specification, and design. Moreover, CT gives developers the means to identify change rapidly and to redesign when the environment changes.

In future articles we intend to:

- Provide more detailed information on the SEC
- To expand our efforts from the media room to other components in the smart home
- In collaboration with other smart home researchers, to integrate these components into a distributed rational agent-based smart home architecture

References

[1] J. W. Priest and J. Sanchez, *Product Development and Design for Manufacturing, A Collaborative Approach to Producibility and Reliability*, New York: Marcel Dekker, 2001.
[2] V. Postrel, "Rising Heat: New Thermostat Designed by Brilliant Morons," *Forbes ASAP*, 1999.
[3] I. Jacobson, G. Booch, and J. Rumbaugh, *The Unified Software Development Process*, Boston: Addison-Wesley, 1999.
[4] W. R. Scott, *Organizations: Chapter 1*, New York: Prentice Hall, 1992.
[5] J. R. Galbraith, *Designing Complex Organizations*, Reading, MA: Addison-Wesley, 1973.
[6] AMX, "An Overview of Home Automation," 2001.
[7] P. Bergstrom, K. Driscoll, and J. Kimball, "Making Home Automation Communications Secure," *IEEE Comp.*, vol. 34, 2001, pp. 50–56.
[8] S. R. Schach, *Object-Oriented and Classical Software Engineering*, 5th ed., New York: McGraw-Hill, 2001.
[9] J. R. Hauser and D. Clausing, "The House of Quality," *Harvard Bus. Rev.*, 1988, pp. 63–73.
[10] M. I. Elboushi and J. S. Sherif, "Object-oriented Software Design Utilizing Quality Function Deployment," *J. Sys. Software*, vol. 38, 1997, pp. 133–43.

Biographies

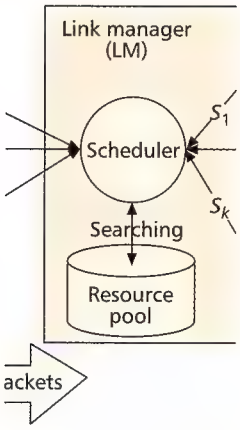
JOHN R DURRETT (john@durrett.org) is an assistant professor of information systems at Texas Tech University. He received his Ph.D. from the University of Texas at Austin in information systems, his M.B.A. and B.A. from West Texas A&M. His research and teaching interests are distributed systems design, elearning, and network security.

LISA BURNELL (l.burnell@tcu.edu) is an assistant professor in the Computer Science Department at Texas Christian University. Her previous experience is in the aerospace and transportation industries. She has a Ph.D. and Master's in computer science and a B.A. in mathematics, all from the University of Texas, Arlington. Her research interests are probabilistic reasoning, decision-theoretic inference, and software engineering methodologies.

JOHN W. PRIEST is a professor of industrial and manufacturing engineering at the University of Texas at Arlington. He is author or co-author of more than 130 technical articles, predominantly on the product development process, design for manufacturing, and technical risk management. Between 1978 and 1999 he has worked on government task forces to improve the product development process including decision support systems and knowledge management. Prior to joining academia, he worked full-time for Rockwell International Communications (now Alcatel), Texas Instruments, and General Motors.

AN ADAPTIVE SNIFF SCHEDULING SCHEME FOR POWER SAVING IN BLUETOOTH

TING-YU LIN AND YU-CHEE TSENG, NATIONAL CHIAO-TUNG UNIVERSITY



Bluetooth is expected to be an important basic constructing component for Smart Homes. In a Smart Home environment, many devices will be portable and battery-operated, making power saving an essential issue. The authors study the problem of managing the low-power sniff mode in Bluetooth.

ABSTRACT

Bluetooth is expected to be an important basic constructing component of smart homes. In a smart home environment, many devices will be portable and battery-operated, making *power saving* an essential issue. In this article we study the problem of managing the low-power *sniff mode* in Bluetooth, where a slave is allowed to be awake only periodically. One challenging problem is how to schedule each slave's sniffing period in a piconet so as to resolve the trade-off between traffic and power-saving requirements, to which we refer as the *sniff-scheduling problem*. We propose an adaptive protocol to dynamically adjust each slave's sniff parameters, with a goal of catching the varying, and even asymmetric, traffic patterns among the master and slaves. Compared to existing works, our work is unique. First, our scheduling considers multiple slaves simultaneously. Existing work only considers one slave, and different slaves are treated independently. Second, our scheduling is more accurate and dynamic in determining the sniff-related parameters based on slaves' traffic patterns. Most work is restricted to a naive exponential adjustment in sniff interval/sniff-attempt window. Third, our proposal includes the placement of sniff-attempt periods of sniffed slaves on the time axis when multiple slaves are involved. This issue is ignored by earlier work. Extensive simulation results are presented. Among many observations, one interesting result is that with proper settings, our protocol can save significant power while achieving higher network throughput than a naive always active round-robin scheme.

INTRODUCTION

Computing and communication anytime, anywhere is a global trend in development today. Ubiquitous computing has been made possible by the advance of wireless communication technology and the availability of many lightweight compact portable computing devices. This area has attracted a lot of attention recently, and various types of network architectures have been proposed, such as wireless LANs, ad hoc networks, sensor networks, and personal area networks.

One emerging environment gaining more and more attention is the *smart home*. The basic idea behind smart homes is to provide various human-friendly services with the goal of facilitating human life. Typical home electronic appliances will not be considered clumsy anymore. Instead, they are capable of coordinating with each other and adapting to surroundings. Such capabilities are achieved by equipping these appliances with embedded computing and communication devices. There are diverse aspects and technologies involved in the development of Smart Homes. One promising technology supported by numerous organizations and companies is Bluetooth. With the design goals of compactness, low cost, and low power, Bluetooth is expected to be a promising basic constructing component of smart homes.

This article focuses on Bluetooth [1], which is characterized by indoor low-power low-complexity short-range radio wireless communications with a frequency-hopping time-division duplex channel model. A master-slaves configuration called a *piconet* is adopted. Readers can refer to [2–4] for more general details of the Bluetooth standard description.

Since low cost is one design goal of Bluetooth, a large number of widespread deployments of Bluetooth are expected. Within home environments, the deployment could consist of various portable devices. One essential issue for almost all kinds of portable devices is *power saving*. Mobile devices have to be supported by batteries, and without power they become useless. Battery power is a limited resource, and it is expected that battery technology is not likely to progress as fast as computing and communication technologies do. Hence, lengthening the lifetime of batteries in portable devices is an important issue. Solutions for power saving can be generally categorized into several approaches.

Transmission power control: In wireless communication, transmission power has a strong impact on bit error rate, transmission rate, and inter-radio interference. These are typically contradicting factors. Power control to reduce interference for ad hoc networks is addressed in [5]. Dynamically adjusting transmission powers of mobile hosts in ad hoc networks to control network topology, known as *topology control*, is addressed in [6]. Increasing network throughput by power adjustment for packet radio networks is addressed in [7].

1536-1284/02/$17.00 © 2002 IEEE

Power-aware routing: Power-aware routing protocols for ad hoc networks are discussed in [8, 9].

Management of low-power modes: More and more wireless devices can provide low-power modes. IEEE 802.11 has a power-saving mode in which a radio only needs to be awake periodically. HiperLAN allows the mobile host, which is in power-saving mode, to define its own active period [10]. As for active hosts, they can save power by turning off their equalizers according to the transmission bit rate. Bluetooth provides three low-power modes: *sniff*, *hold*, and *park* [1].

We study the management of low-power sniff mode in Bluetooth to conserve power; thus, this falls into the third category above. In sniff mode, a slave's listening activity is reduced. Slaves only listen in specified time slots regularly spaced by sniff intervals. One challenging problem is how to schedule each slave's sniffing period in a piconet to balance the trade-off between traffic and power-saving requirements, to which we refer as the *sniff scheduling problem*.

In this article, an adaptive sniff scheduling protocol is proposed to periodically adjust the sniff parameters. An *evaluator* is used by each master and its slaves to determine its traffic pattern and sniff-related parameters. A *scheduler* is deployed on the master's side to schedule each slave's sniffing period. Since each slave's sniffing period can be regarded as an infinite sequence of time slots and multiple slaves are considered in this article, we propose a concept called *resource pool* (RP) to manage the available/occupied time slots in the piconet. The master periodically checks the needs of its slaves by running the evaluator, and allocates suitable slot resources from the RP for them. On the other hand, slaves can exercise their own evaluators and issue requests for slot resources as well. Two scheduler policies are proposed: *Longest Sniff Interval First* (LSIF) and *Shortest Sniff Interval First* (SSIF). Simulation results are presented to verify the effectiveness of our protocol.

Compared to existing work, our work is unique in the following senses. First, our scheduling scheme considers multiple slaves simultaneously. Existing work only considers one slave, and different slaves are treated independently. Second, our scheduling is more accurate and dynamic in determining the sniff-related parameters based on slaves' traffic patterns. Most work is restricted to a naive exponential adjustment in sniff interval/sniff-attempt window. Third, our proposal includes the placement of sniff-attempt periods of sniffed slaves on the time axis when multiple slaves are involved. This issue is ignored by earlier work.

Related work includes [11–14]. In [12, 14], the polling priorities of slaves are determined based on their traffic loads; however, how to combine this with the sniff mode is not addressed. In [13] it is proposed to dynamically adjust the sniff parameters according to a slave's slot utilization. In [11] a learning function is proposed to determine the sniff interval. However, the sniff interval is adjusted on a per-interval basis, and thus the overhead of control messages might be pretty high. Both [11, 13] suffer the problems that only one single slave is considered in an independent way, and the placement of sniff-attempt windows is not addressed.

The rest of this article is organized as follows. Preliminaries are given in the next section, followed by our sniff scheduling protocol. We propose two policies for our scheduler. Simulation results are then provided. Finally, we draw conclusions.

Preliminaries

Bluetooth Technology Review

Bluetooth is a master-driven time-division duplex short-range radio wireless system. The smallest network unit is called a *piconet*, and has a master-slaves configuration. A time slot in Bluetooth is 625 μs. The master sends data to slaves in even-numbered slots, while slaves send data to the master in odd-numbered slots. A slave only transmits packets after the master polls or sends data to it.

According to the Bluetooth protocol stack [1], on top of RF is the Bluetooth baseband, which controls use of the radio. Four important operational modes are supported by the baseband: *active*, *sniff*, *hold*, and *park*. Active mode is most energy-consuming, where a Bluetooth unit is turned on most of the time. Sniff mode allows a slave to go to sleep and only wake up at a specific time. In hold mode, a slave can temporarily suspend supporting data packets on the current channel; the capacity can be made free for other things, such as scanning, paging, and inquiring. While in hold mode, a unit can also attend other piconets. Prior to entering hold mode, an agreement should be reached between the master and slave on hold duration. When a slave does not want to participate in the piconet channel, but still wants to remain synchronized, it can enter park mode. The parked slave has to wake up regularly to listen to the channel, to stay synchronized or check broadcast messages.

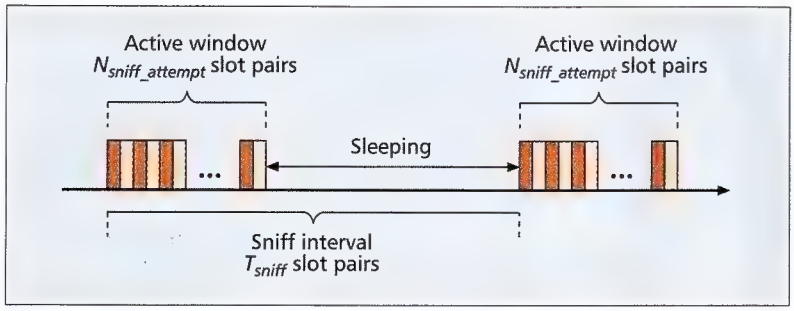
On top of the baseband is the link manager (LM), which is responsible for link configuration and control, security functions, and power management. The corresponding protocol is called Link Manager Protocol (LMP). The Logical Link Control and Adaptation Protocol (L2CAP) provides connection-oriented and connectionless datagram services to upper-layer protocols. Two major functionalities of L2CAP are protocol multiplexing, and segmentation and reassembly (SAR).

The Service Discovery Protocol (SDP) defines the means for users to discover which services are available in their neighborhood and the characteristics of these services. The RFCOMM protocol provides emulation of serial ports over L2CAP to support many legacy applications based on serial ports over Bluetooth without any modifications. Up to 60 serial ports can be emulated.

The Low-Power Sniff Mode

Our main focus, the power-saving issue of Bluetooth, is discussed in more detail in this section. In active mode, the Bluetooth unit is turned on most of the time to participate in send/receive activities. An active slave listens in even slots

According to the Bluetooth protocol stack, on top of RF is the Bluetooth Baseband, which controls the use of the radio. Four important operational modes are supported by the baseband: active, sniff, hold, and park.



■ Figure 1. *Sniff interval and active window of Bluetooth (darkened parts are even slots).*

for packets. If the slave is not addressed in the current packet, it may sleep until the next even slot the master transmits. From the type indication of the packet, the slave can derive the number of slots to be used by the master to transmit the current packet. The addressed slave will reply in the next odd slot after the master's transmission.

In the sniff mode, the slave's listening activities are reduced to save energy. For a sniffed slave, the time slots when the master can communicate to that slave are limited to some specific time slots. These so-called *sniff-attempt slots* arrive periodically, as illustrated in Fig. 1. In the Bluetooth specification, there are three parameters specified for such sniff activity: T_{sniff}, $N_{sniff\_attempt}$, and $N_{sniff\_timeout}$. Since Bluetooth specification separates even and odd slots for the master and slave transmissions, the values of these sniff parameters are all based on *slot pairs* (one even plus one odd slot). In every T_{sniff} slot pair, the slave will wake up to listen to the master for $N_{sniff\_attempt}$ consecutive (even) slots for possible packets destined to it. After every reception of a packet with a matching address, the slave continues listening for $N_{sniff\_timeout}$ more slots or for the remaining of the $N_{sniff\_attempt}$ slot pairs, whichever is greater. In this article, we call T_{sniff} the *sniff interval* and $N_{sniff\_attempt}$ the *active window*.

The control packets exchanged between two communicating LMs via LMP are termed LMP_PDUs. There are four LMP_PDUs involved in sniff mode management: LMP_sniff_req, LMP_accepted, LMP_not_accepted, and LMP_unsniff_req. These PDUs are for making/rejecting/accepting requests and returning to normal active mode. To enter sniff mode, an LMP_sniff_req request packet can be initiated by either a master or a slave carrying the proposed parameters. Upon receipt of the request, the receiver side can negotiate with the other side on the related sniff parameters by issuing another LMP_sniff_req request packet carrying the suggested parameters. If an agreement can be seen, an LMP_accepted packet is used to place the slave into sniff mode. Otherwise, an LMP_not_accepted packet is returned with a reason code for rejection. Also, note that the sniff parameters can be negotiated on a per-master-slave basis.

Sniff mode can be ended by sending an LMP_unsniff_req packet. The counterpart must reply with an LMP_accepted packet. If this is requested by the slave, it will enter active mode after receiving LMP_accepted. If this is requested by the master, the slave will enter active mode immediately after receiving LMP_unsniff_req.

Problem Statement

Although the operations of sniff mode are given in the Bluetooth specification, how to determine the sniff-related parameters is left as an open issue for the designers. One should dynamically determine a proper set of sniff parameters for a slave based on its traffic pattern to save as much of its power as possible without incurring too much delay in packet delivery. Furthermore, multiple slaves could be in sniff mode at the same time. How to schedule their active windows on the time axis is a challenging problem since these windows arrive periodically and may extend, conceptually, to infinity on the time axis. Overlapping of active windows is allowed but undesirable. Collectively, we call this the *sniff scheduling problem*.

Finally, we are aware of the possibility of using hold or park mode for power saving. However, this article only considers sniff mode because it can serve various types of traffic with power saving in mind.

The Sniff Scheduling Protocol

In this section we propose a protocol to exploit the low-power sniff mode of Bluetooth. Because of the master-driven centrally controlled architecture of Bluetooth, we will maintain an RP on the master's side. The sniff parameters of each slave will be adjusted dynamically based on many factors such as the slave's traffic load, current backlog, previous utilization of sniff slots, and availability of the RP. The ultimate goal is to save slaves' power while keeping packet delays as small as possible. Note that because of Bluetooth's separation of even and odd slots, all calculations refer to slot pairs unless stated otherwise.

Figure 2 shows the architecture of our sniff scheduling protocol in a piconet with K active slaves, $1 \le K \le 7$. On the master side, there are three main entities: evaluator, scheduler, and RP. The master periodically runs the evaluator for each slave k, $1 \le k \le K$, to evaluate its condition. If necessary, a value S_k, which is used to reflect the estimated traffic load of slave k, is generated and fed into the scheduler to readjust this slave's sniff parameters. The scheduler then searches the RP to schedule a new set of sniff parameters for the slave.

On the other hand, slaves also run their own evaluators periodically. These distributed evaluators will pass their desired S_k values to the master via an LMP_sniff_req packet. The scheduler then searches the RP, and arranges new sniff parameters for them. In our protocol, when the scheduler is unable to find suitable sniff parameters for a slave, an LMP_unsniff_req packet will be issued to invite it into active mode. This usually happens when the traffic load of the slave is quite large. When the master finds it possible to allocate a proper sniff scheduling for the slave, it may bring the slave back to sniff mode again.

Since the low-power modes of Bluetooth are

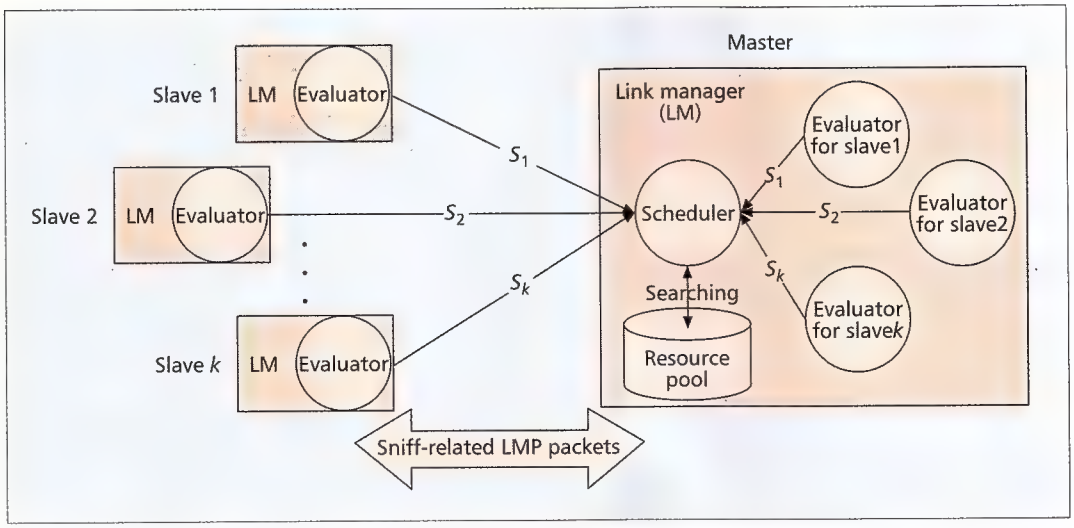


Figure 2. *The architecture of our sniff scheduling protocol.*

managed by the link manager (LM), our protocol should reside in the LM layer of each Bluetooth unit, monitoring the backlogs of the lower baseband buffers and issuing proper sniff scheduling packets. In this article we follow the same assumption as in [12]: the master keeps separate buffers in the baseband layer for its slaves. Each buffer queues the data dedicated to the corresponding slave. Below, we discuss our design in more detail.

The Evaluator

The purpose of the evaluator is to measure how efficiently or inefficiently a slave uses the sniff-attempt slots assigned to it and, if necessary, to trigger the scheduler to readjust its sniff parameters.

The fundamental parameters are explained below. First, we define (T_k, N_k, O_k) as the current sniff parameters (sniff interval, active window size, and offset, respectively) associated with slave k. For each slave k, we have to measure its U_k. This is a ratio between (including) 0 and 1, indicating how many slot pairs of the sniff-attempt slots are used effectively for real data communication under the current setting for slave k. Also, we use B_k to denote the buffer backlog for slave k (i.e., the number of packets currently queued in the local baseband buffer). By U_k and B_k, we derive a weighted value W_k to measure the current requirement of slave k:

$$W_k = \alpha U_k + (1 - \alpha)B_k/B_{max}, \quad (1)$$

where B_{max} is the maximum buffer space, and α is a constant between 0 and 1 to differentiate the importance of U_k and B_k. The resulting W_k will be tested against the condition $r_{lb} < W_k < r_{ub}$, where r_{lb} and r_{ub} are predefined tolerable lower bound and upper bound, respectively, of W_k. If this condition is violated, the master will be triggered to readjust slave k's current sniff parameters; otherwise, no readjustment is needed.

Based on W_k, our objective is to determine the desired slot occupancy S_k of slave k, which represents the expected ratio of the new N_k to the new T_k:

$$S_k = (N_k/T_k) \times W_k/\delta. \quad (2)$$

Intuitively, N_k/T_k is slave k's current slot occupancy. Multiplying this ratio by W_k gives the slot occupancy ratio expected to be assigned to this slave. The factor δ is a positive constant below 1 to enlarge the expected ratio to tolerate a certain level of inaccuracy in our estimation.

Note that a minor logic flaw we intentionally omit in the above discussion (for ease of presentation) is that when the slave is in active mode, it has no sniff parameters, so the ratio N_k/T_k becomes meaningless. In this case we simply replace this ratio by the recent slot occupancy of slave k (with all slots as the denominator). The rest is all the same.

The Resource Pool

The available sniff-attempt slots that can be allocated to slaves are managed by the RP on the master side. One may regard the sniff-attempt slots of a slave as a periodical infinite sequence. However, we need an efficient finite data structure for representation.

In this section we propose to use a two-dimensional $d_1 \times d_2$ matrix M for the representation. The basic idea is as follows. We group (infinite) slot pairs appearing with period $d_1 \cdot d_2$ into one set. For $p = 0..d_1 \cdot d_2 - 1$, define

$$G_p = \{p + q \cdot d_1 \cdot d_2 \mid q \text{ is any non-negative integer}\}.$$

Each entry in matrix M is used to represent the availability of one such slot group, so we define, for $i = 0..d_1 - 1, j = 0..d_2 - 1$,

$$M[i,j] = \begin{cases} 0 \text{ if } G_{i \cdot d_2 + j} \text{ is free} \\ 1 \text{ if } G_{i \cdot d_2 + j} \text{ is busy} \end{cases}. \quad (3)$$

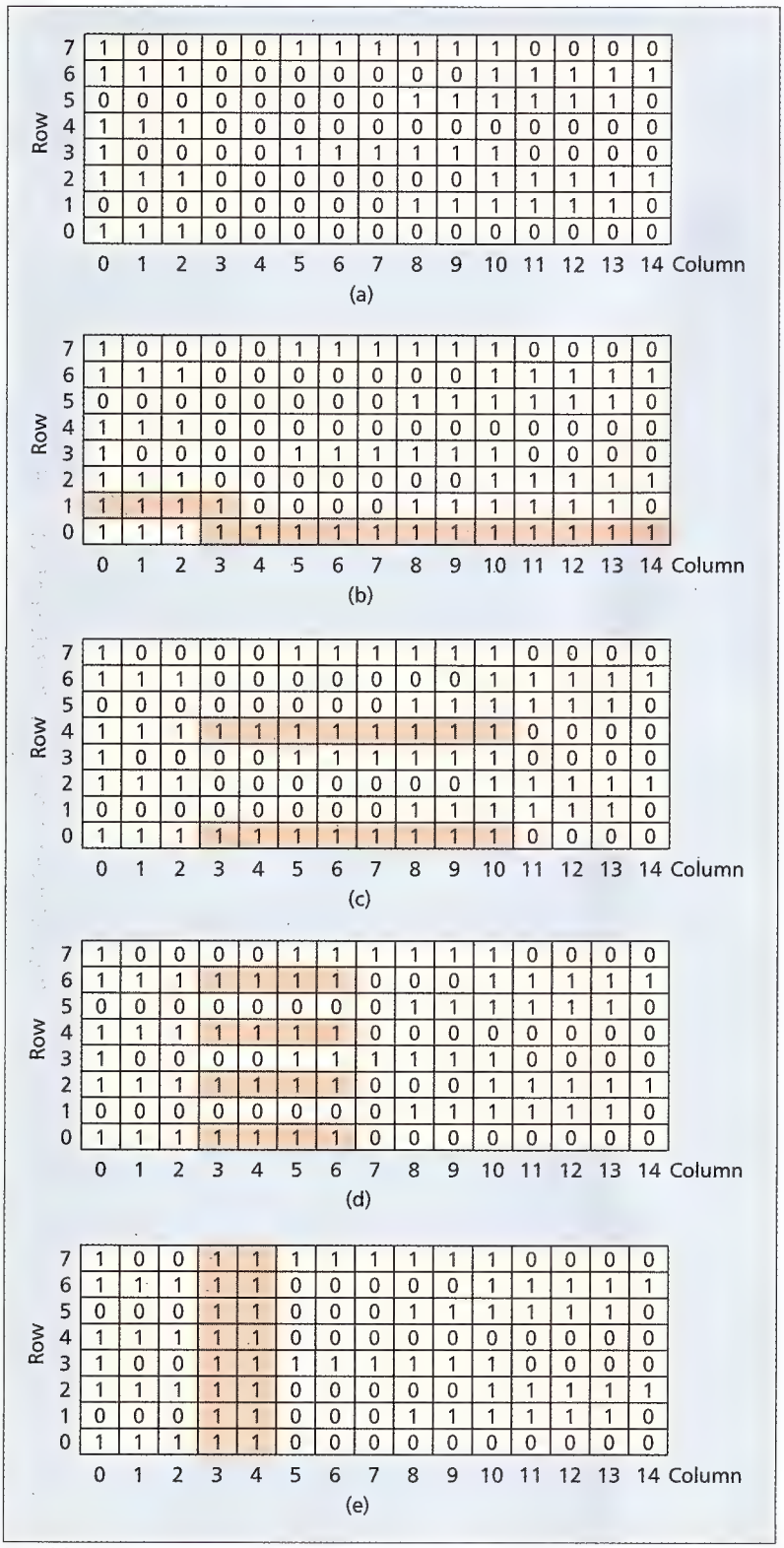
Note that variables d_1 and d_2 are adjustable parameters. The value of $d_1 \cdot d_2$ should be large enough to capture the behavior of those slaves that have very low traffic load and would like to spend very low energy on sniff-attempts. Also note that $d_1 \cdot d_2$ indicates the maximum allowed sniff interval. To facilitate exponential adjustment of sniff intervals, we configure d_1 to be power of 2, denoted 2^u, where u is an non-negative integer. Besides, d_2 is replaced with T, which indicates the minimum allowable sniff interval.

Since the low-power modes of Bluetooth are managed by the link manager, our protocol should reside in the LM layer of each Bluetooth unit, monitoring the backlogs of the lower baseband buffers and issuing proper sniff-scheduling packets.

Different values of u and T will provide different levels of flexibility, as shown later. Table 1 plots several examples of M with size 8×15.

The 2D matrix M can provide us much flexibility in managing periodical time slots. We can manipulate both sniff intervals and active windows of sniffing slaves easily. Two groups that are adjacent in M can be framed together to double the active window. Two groups spaced by a certain distance in the matrix can be grouped together too to divide the sniff interval by half. This can be extended to the combination of more groups easily. Reversibly, we may decrease the active window or increase the sniff interval of a slave to reduce its power consumption by partitioning groups. For example, given the state in Table 1a, if a slave's packet mean arrival rate is 16/120, we show four possible ways (b, c, d, e) to arrange the slave's sniff-attempt slots. Table 1b indicates that the slave is awake every 120 slot pairs, each time lasting for 16 slot pairs. Table 3c shows that it is awake every 60 slot pairs, each time lasting for 8 slot pairs. Table 1d means it is awake every 30 slot pairs, each time lasting for 4 slot pairs. Table 1e means it is awake every 15 slots pairs, each time lasting for 2 slot pairs.

With such a formulation, the resource management problem becomes one of allocating proper entries in matrix M. Later, we will propose two searching policies for this problem.



■ Table 1. *Example: Initial state of an 8 × 15 matrix* M *and feasible assignments in matrix* M *(dark) for a slot occupancy of 16/120.*

LMP_PDU Flows

In the Bluetooth specification, both master and slaves can initiate a sniffing request. In our protocol, we also allow a sniffing request to be master- or slave-activated. This section discusses the related LMP_PDU exchanges. Recall the calculation of W_k for slave k. All the following discussion is triggered by violating the constraint $r_{lb} < W_k < r_{ub}$.

Figure 3 shows four possible master-activated scenarios. The first one demonstrates the master proposing a new set of sniff parameters (O_k', T_k', N_k') for slave k. In response, the slave runs its evaluator to determine its local S_k. If the assigned slot occupancy derived from N_k'/T_k' is not less than S_k, an LMP_accepted can be returned.

The second scenario demonstrates that slave k disagrees on the assigned parameter, so an LMP_sniff_req is returned. However, note that since slave k does not know the status of the RP, we actually intend to return its estimated S_k to the scheduler for an arrangement. Since S_k is a ratio between 0 and 1, we simply use the two fields T_{sniff} and $N_{sniff\_attempt}$ in LMP_sniff_req to carry two values, T_k'' and N_k'', respectively, such that $N_k''/T_k'' \approx S_k$. The scheduler then tries to allocate a new set of sniff parameters based on N_k''/T_k'' for slave k. In response, the slave issues an LMP_accepted.

The third scenario is similar to the second one, but the master fails to allocate a large enough active window from its RP (probably because matrix M is too crowded or fragmented). In this case, the master will request the slave to unsniff itself.

The fourth scenario is where, from the estimated S_k, the master directly finds it has difficulty allocating a large enough active window from its RP, so an unsniff request is directly sent to the slave.

Figure 4 shows slave-activated negotiation. This is similar to the aforementioned second and third scenarios, but it is triggered by finding violation of the condition $r_{lb} < W_k < r_{ub}$ on the slave's side. Again, since the slave does not know

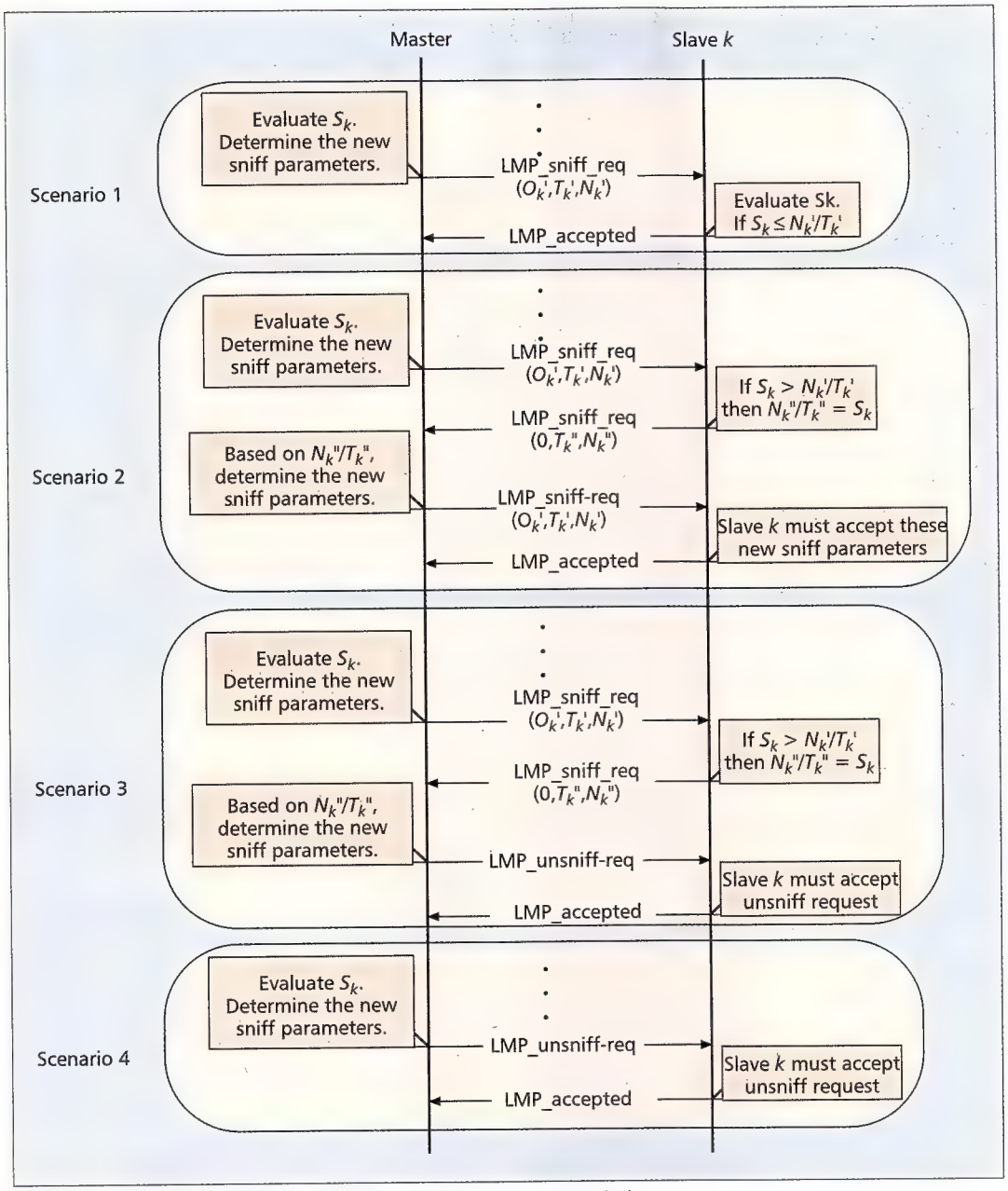


Figure 3. *Scenarios of master-activated sniff parameter negotiation.*

the status of the RP, we use the two fields T_{sniff} and $N_{sniff\_attempt}$ in LMP_sniff_req to convey the desired S_k to the master. It then follows with an LMP_sniff_req containing the new sniff parameters or an LMP_unsniff_req packet, depending on the crowdedness of the RP.

Scheduling Policies for the Resource Pool

The job of the scheduler is to take an input S_k and determine a suitable set of sniff parameters (denoted O_k', T_k', and N_k' below) for slave k. These parameters in fact represent a set of slot groups in matrix M. Before making the allocation, the old occupancy by this slave on M should be released, which is an easy job. In the following, we propose two policies for searching M, named *LSIF* and *SSIF*.

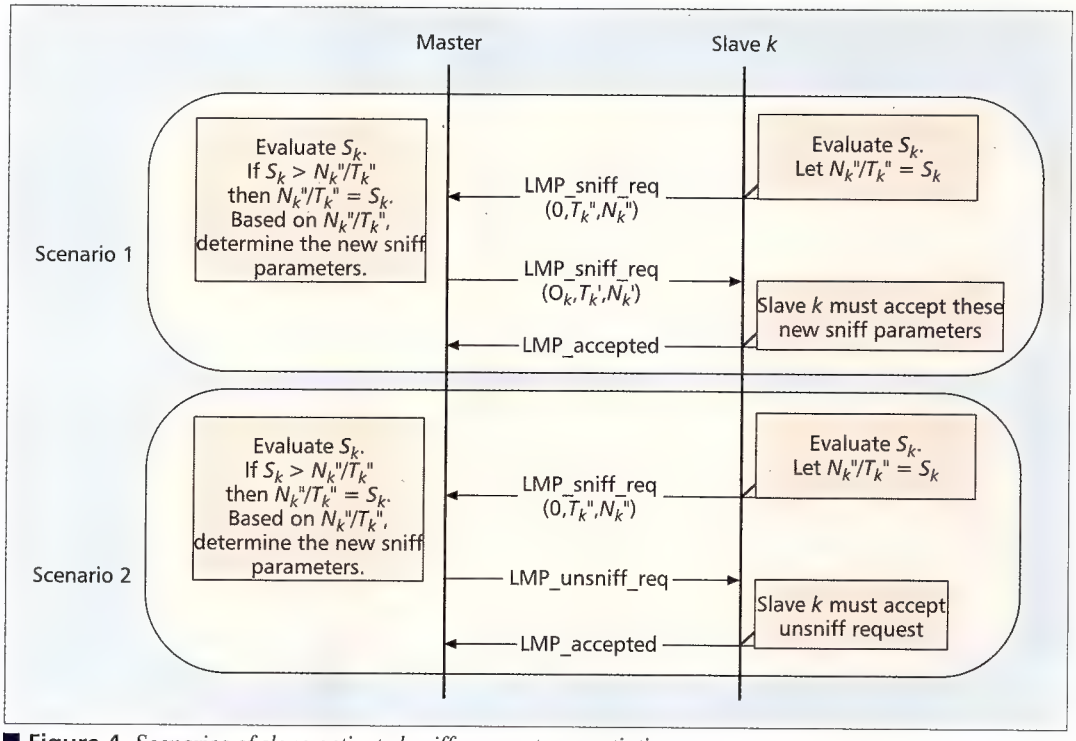
Longest Sniff Interval First

In the LSIF policy, we search matrix M starting from the longest sniff interval, which is $2^u \cdot T$. If the search fails, we divide the interval by 2, which is $2^{u-1} \cdot T$, and do the search again. If the search also fails, we further use the interval $2^{u-2} \cdot T$ to do the search. The search stops once a satisfactory set of slot groups is found. This is repeated until the shortest interval T is tried, in which case we will bring the slave into active mode, as discussed earlier.

Below, we discuss the detailed steps when we search matrix M with a sniff interval $2^p \cdot T$, where $0 \le p \le u$. Since the matrix is of size $2^u \times T$, we will "fold" M into a smaller matrix of size $2^p \times T$. Specifically, we partition M horizontally and evenly into 2^{u-p} pieces, each of size $2^p \times T$. Then we fold all pieces together by executing a bitwise OR operator. The meaning of OR is to

The available sniff-attempt slots that can be allocated to slaves are managed by the resource pool at the master side. One may regard the sniff-attempt slots of a slave as a periodical, infinite sequence. However, we need an efficient, finite data structure for the representation.

We have developed a simulator to verify the effectiveness of our protocol. The goal is to observe the interaction between the two seemingly contradicting factors: network throughput and power consumption.



■ **Figure 4.** *Scenarios of slave-activated sniff parameter negotiation.*

ensure that each piece of submatrices has a free entry. Let the folded matrix be M'. We then search M' sequentially for possible existence of q consecutive free entries (with value 0), where

$$q = \left\lfloor \frac{S_k \cdot 2^u \cdot T}{2^{u-p}} \right\rfloor = \left\lfloor S_k \cdot 2^p \cdot T \right\rfloor.$$

The reason we adopt a floor function instead of a ceiling function here is that the desired slot occupancy has been enlarged in our calculation (by dividing by a $\delta < 1$). Once this search succeeds, we can return $T_k' = 2^p \cdot T$, $N_k' = q$, and $O_k' = i \cdot T + j$, where O_k' is the offset indicating the starting point, say $M'[i, j]$, of the q consecutive free entries in M'.

An example is in Table 2, given $T = 15$, $u = 3$, $r_{lb} = 0.2$, $r_{ub} = 0.8$, and $\delta = 0.8$. Assuming that there are $K = 5$ slaves, in the beginning round 0 all slaves share 1/5 of slot groups. In round 1, the estimated W_2 of slave 2 decrease to 0.18. So we release its occupancy on M and allocate a new space for it. In the table, the ratio q/t means that the target number of 0's on M' is q and the searched sniff interval is t. For example, in round 1 we succeed in ratio 2/60 (underlined), which means we find two consecutive 0s when the searched sniff interval is 60. Similarly, in rounds 2 and 3 we succeed in ratios 3/120 and 6/30, respectively.

Shortest Sniff Interval First

The SSIF policy only differs from the LSIF policy in that it searches starting from smaller sniff intervals and gradually increasing the searched interval. Specifically, we will start from the shortest sniff interval T. If the search fails, we double the interval and repeat the search, until the longest interval $2^u \cdot T$ is tried. The intuition is that although the slot occupancy may remain the same, with a smaller sniff interval the buffers for this slave may experience less chance of overflow. Hence, SSIF has potential to improve network throughput. However, the cost is put on the energy, since the slave needs to wake up more frequently. An example of this policy is in Table 3 (with all inputs the same as in Table 2).

One minor detail we intentionally omitted in the above discussion is that when the slave's traffic load is very low, it is possible that the value of q is always less than 1 throughout the searching. In this case, we simply take the sniff interval $2^p \cdot T$ such that $S_k \cdot 2^p \cdot T$ is closest to 1 to do the search again by enforcing $q = 1$.

Simulation Results

We have developed a simulator to verify the effectiveness of our protocol. The goal is to observe the interaction between the two seemingly contradicting factors *network throughput* and *power consumption*. We simulated a single piconet with six Bluetooth units (one master and five slaves). The programming environment was GNU C++ on UNIX SunOS 5.7. No mobility was modeled (i.e., no device joining or leaving the piconet during the simulation process). The power consumption of the master was not a concern since we assume it has plug-in electricity. Each slave might switch between active and sniff modes, and we didn't consider other modes, such as hold and park. When switching modes or changing sniff parameters, hosts send control packets as described in Figs. 3 and 4. For each slave there is a separate buffer queue on the master side. Physical transmission problems such as fading and interference were not taken into account.

Our power model is derived based on experi-

ences in Lucent WaveLAN cards and Bluetooth [13], which is summarized below. It takes half a unit of power for a Bluetooth device to receive a one-slot packet, and one unit to transmit a one-slot packet. Voice traffic is not simulated (but we can simply reserve 1s in matrix M spaced by regular distances to model such traffic). When a slave hears packets dedicated to others, a lower amount of power is required since it can turn off its receiver after monitoring the packet header. In this case, it only consumes 1/6 of receiving power, which is 0.083 units. Bluetooth also defines some short packets, such as ACK and NULL (to respond to a poll with no data). We assume the power consumption for delivering such packets to be 1/6 of the transmission power, which is 0.167 units.

In addition to our LSIF and SSIF policies, three other approaches are simulated for comparison. The first one is called Always Active (AA), where we enforce all slaves to always stay in active mode. The master polls its slaves using a round-robin policy. Note that, while naive, AA is used here only as a reference point. The second approach is called Always-Sniff-with-Varying-Sniff-Interval (AS_VSI), where we enforce slaves to always stay in sniff mode, but the active window must remain constant. When a slave's slot utilization $\leq r_{lb}$, its sniff interval will be doubled. On the other hand, its sniff interval will be cut in half if slot utilization $\geq r_{ub}$. The smallest sniff interval is T, while the largest possible is $2^u \cdot T$. The third approach is called Always-Sniff-with-Varying-Active-Window (AS_VAW), where we enforce the sniff interval to be constant. The active window will be doubled when slot utilization $\geq r_{ub}$, but cut in half when $\leq r_{lb}$. The lower and upper bounds of active window size are 1 and $T/5$, respectively.

Below, we divide our presentation into two parts. The next section shows the results under fixed traffic patterns, the following section under varying traffic patterns. The latter is intended to model real system traffic and demonstrate the flexibility of our protocol in catching such dynamics. Through the presentation we shall provide a guideline for choosing proper parameters for our protocol. In our simulation we always adopt r_{lb} = 0.3 and r_{ub} = 0.7. Each simulation run lasts for 100,000 slot pairs.

Round0 $N_1/T_1=N_2/T_2=N_3/T_3=N_4/T_4=N_5/T_5=3/15$

S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5

Round1 $W_2=0.18<r_{lb} \Rightarrow S_2=[(3/15)*0.18]/0.8=0.045$
$0.045=5/120=\underline{2/60}=1/30=0/15 \Rightarrow (O_k',T_k',N_k')=3,60,2)$

S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2		S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2		S3	S3	S3	S4	S4	S4	S5	S5	S5

Round2 $W_3=0.11<r_{lb} \Rightarrow S_3=[(3/15)*0.11]/0.8=0.028$
$0.028=\underline{3/120}=1/60=0/30=0/15 \Rightarrow (O_k',T_k',N_k')=(5,120,3)$

S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2					S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S3	S3	S3		S4	S4	S4	S5	S5	S5

Round3 $W_4=0.9>r_{lb} \Rightarrow S4=[(3/15)*0.9]/0.8=0.23$
$0.23=27/120=13/60=\underline{6/30}=3/15 \Rightarrow (O_k',T_k',N_k')=(18,30,6)$

S1	S1	S1	S4	S4	S4	S4	S4	S4				S5	S5	S5
S1	S1	S1										S5	S5	S5
S1	S1	S1	S4	S4	S4	S4	S4	S4				S5	S5	S5
S1	S1	S1	S2	S2								S5	S5	S5
S1	S1	S1	S4	S4	S4	S4	S4	S4				S5	S5	S5
S1	S1	S1										S5	S5	S5
S1	S1	S1	S4	S4	S4	S4	S4	S4				S5	S5	S5
S1	S1	S1	S2	S2	S3	S3	S3					S5	S5	S5

■ Table 2. *Searching example of LSIF.*

Fixed Traffic Load

Here we assume a Poisson process with packet arrival rate λ = 0.2 (packets per time slot) for each buffer (on both the master and slave sides). Figure 5 illustrates the power consumption and network throughput against buffer size B_{max}. We observe that with enough buffer space ($\geq$ 50), all five approaches can achieve high throughput close to 0.99. This is because fixed traffic loads are easy to catch. Compared to AA, the other four schemes consume significantly less power.

These observations motivate us to derive the following simulations, where traffic is nonuniform, to reveal the strength of our proposals.

Varying Traffic Load

Here we adopt a variable traffic model similar to that proposed in [14]. Packets still arrive by the Poisson process, but at different rates. Two types of traffic patterns will be explored. The first is denoted TypeI(λ_M–λ_S), which means the arrival rate on the master's side is λ_M, and that on the slave's side is λ_S. The second is denoted TypeII(λ_A–λ_B), which means there are two kinds of arrival rates, λ_A and λ_B, for both the master and the slave. The master and the slave change states between rates λ_A and λ_B independently, and the transition probability from one rate to the other is 0.01 in both directions.

We first investigate the effect of weight α on our LSIF and SSIF schemes. Figure 6a plots the network throughput against α, where five slaves with traffic patterns TypeI(0.2–0.2), TypeI(0.19–0.01), TypeI(0.01–0.19), TypeII(0.19–0.01), and TypeII(0.19–0.01) are simulated. It indicates that

the throughput can be consistently improved as α increases, until α ≤ 0.7. At α = 0.0, LSIF and SSIF give the worst throughput of all. The reason is that the evaluating metric is all based on buffer information when α = 0.0, which is unfair. As a result, the assigned sniff-attempt slots are unable to handle future traffic well, thus degrading performance. After α > 0.7, the throughput starts to degrade as α grows. However, even when α = 1.0, our LSIF and SSIF still outperform the other three schemes. Thus, a value between 0.6 and 0.7 for α could be the best choice.

The above simulation in fact reveals two interesting phenomena. First, the slot utilization factor alone cannot predict the traffic well. One should take both slot utilization and buffer backlog information to predict future traffic. Second, and much to our surprise, our LSIF and SSIF schemes can even provide better network throughput than a naive AA round-robin scheme as α is properly set. The reason is that, in the AA case, the master wastes much time polling slaves with no backlogs, resulting in reduced throughput. This indicates a prospective direction in which one can save power and improve network throughput at the same time (which are contradicting factors by intuition).

Figure 6b illustrates the impact of weight δ on LSIF and SSIF, with α = 0.7 and the same traffic pattern as above. It shows that before δ ≤ 0.6, the throughput can be improved as δ grows. Recall that δ is a factor to enlarge the expected sniff-attempt slots to tolerate a certain level of inaccuracy in our estimation. With a too small value (e.g., δ = 0.1), LSIF and SSIF will degenerate into the AA scheme, since slaves can hardly obtain such a large slot occupancy. As a result, all slaves may remain active most of the time. This also violates our goal of conserving power. After δ > 0.6, the throughput starts to decline as δ increases. After δ > 0.8, our throughput falls behind that of the AA scheme. This is because our prediction is too conservative to catch the dynamics of variable traffic. Thus, a reasonable value for δ would be between 0.5. and 0.7.

Also with the same traffic pattern, in Fig. 7a, we study network throughput with different values of u, which controls the largest possible sniff interval. For both LSIF and SSIF, it shows that the throughput remains at around the same level when u = 0, 1 and 2, but decreases as u is larger. With u = 0, there is only one sniff interval available, which is T. With u = 1 and 2, there are two and three sniff intervals available, respectively. A larger sniff interval means that the slave needs to switch modes less frequently, which is more favorable. Based on these considerations, one may choose a proper value of u to use.

One may note that in the previous simulation, the advantage of using our 2D matrix M is not well justified. When u = 0, M degenerates to a 1D matrix; so why one should need a 2D M remains a question. The reason is that we have adopted T = 100. Since the lowest traffic load we may inject for each entity is 0.01, a sniff interval of T = 100 and active window of 1 can properly catch such traffic without much waste. To justify this point, we have simulated hosts with very low traffic loads. We have conducted another experiment with five slaves having the following traffic patterns: TypeI(0.2–0.2), TypeI(0.001–0.001), TypeI(0.002–0.002), TypeII(0.002–0.005), and TypeII(0.002–0.005). The result is in Fig. 7b. It indicates that the throughput can be improved as u increases, until $u \le 4$. Before $u < 2$, the throughput of LSIF and SSIF is worse than AA. This phenomenon is due to slot waste caused by sniff intervals that are too short. (For example, to handle a low arrival rate of 0.002, the scheduler should reserve one out of every 500 slots in average. When u is set too small, say $u < 2$, the scheduler will reserve too much resource for such slaves. Therefore, both network throughput and power consumption might get hurt.) As mentioned earlier, the value of $2^u \cdot T$ should be large enough to capture the behavior of those slaves that have very low

Round0 $N_1/T_1=N_2/T_2=N_3/T_3=N_4/T_4=N_5/T_5=3/15$

S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S5	S5	S5

Round1 $W2=0.18<r_{lb} \Rightarrow S2=[(3/15)*0.18]/0.8=0.045$
$0.045=5/120=2/60=\underline{1/30}=0/15 \Rightarrow (O_k',T_k',N_k')=(3,30,1)$

S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2			S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2			S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2			S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1				S3	S3	S3	S4	S4	S4	S5	S5	S5
S1	S1	S1	S2			S3	S3	S3	S4	S4	S4	S5	S5	S5

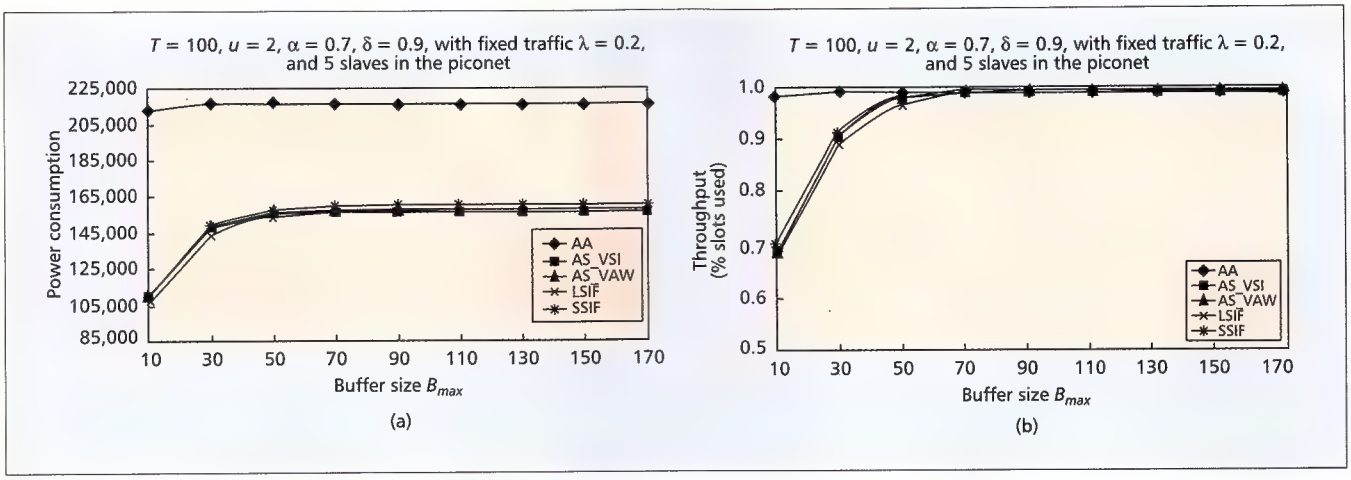
Round2 $W_3=0.11<r_{lb} \Rightarrow S_3=[(3/15)*0.11]/0.8=0.028$
$0.028=3/120=\underline{1/60}=0/30=0/15 \Rightarrow (O_k',T_k',N_k')=(4,60,1)$

S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1	S2						S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S3					S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1	S2						S4	S4	S4	S5	S5	S5
S1	S1	S1							S4	S4	S4	S5	S5	S5
S1	S1	S1	S2	S3					S4	S4	S4	S5	S5	S5

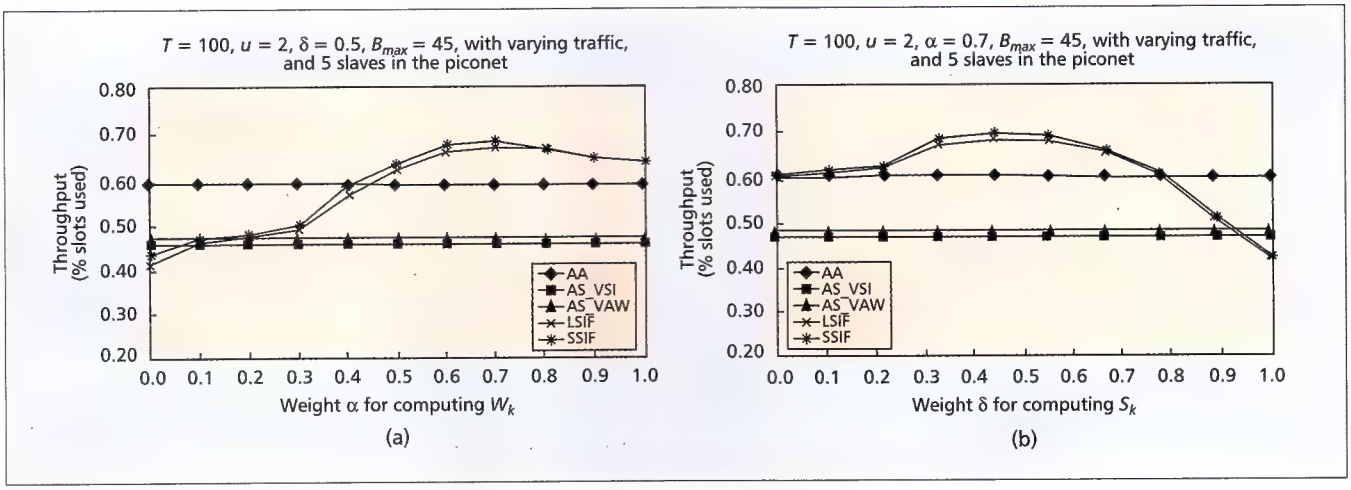
Round3 $W_4=0.9>r_{lb} \Rightarrow S4=[(3/15)*0.9]/0.8=0.23$
$0.23=27/120=13/60=6/30=\underline{3/15} \Rightarrow (O_k',T_k',N_k')=(5,15,3)$

S1	S1	S1			S4	S4	S4					S5	S5	S5
S1	S1	S1	S2		S4	S4	S4					S5	S5	S5
S1	S1	S1			S4	S4	S4					S5	S5	S5
S1	S1	S1	S2	S3	S4	S4	S4					S5	S5	S5
S1	S1	S1			S4	S4	S4					S5	S5	S5
S1	S1	S1	S2		S4	S4	S4					S5	S5	S5
S1	S1	S1			S4	S4	S4					S5	S5	S5
S1	S1	S1	S2	S3	S4	S4	S4					S5	S5	S5

■ Table 3. *Searching example of SSIF.*



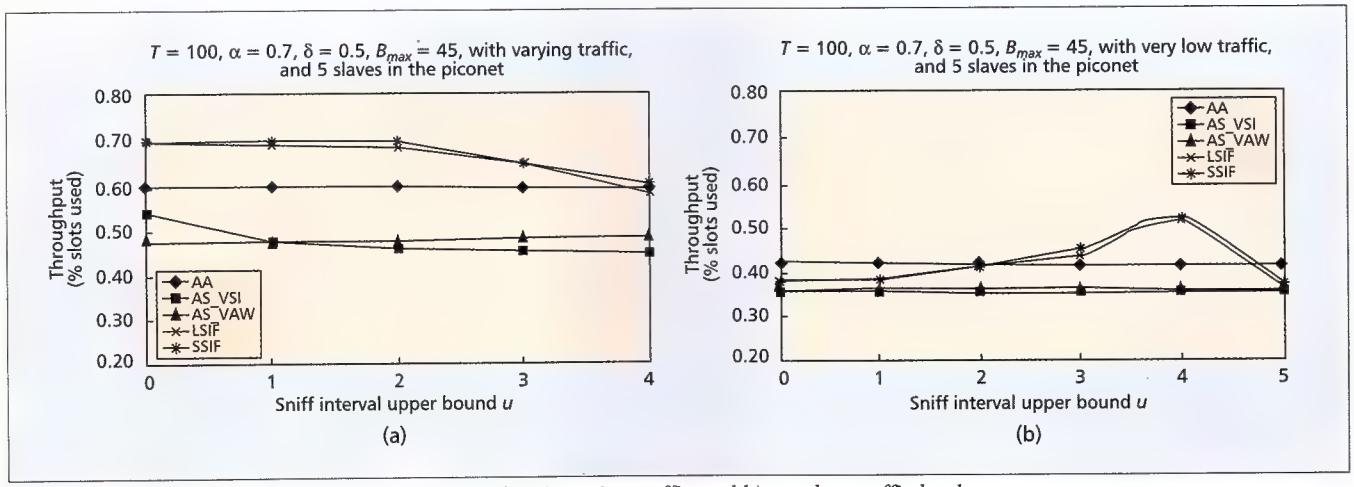
■ Figure 5. *The effect of* B_{max} *under fixed traffic load.*



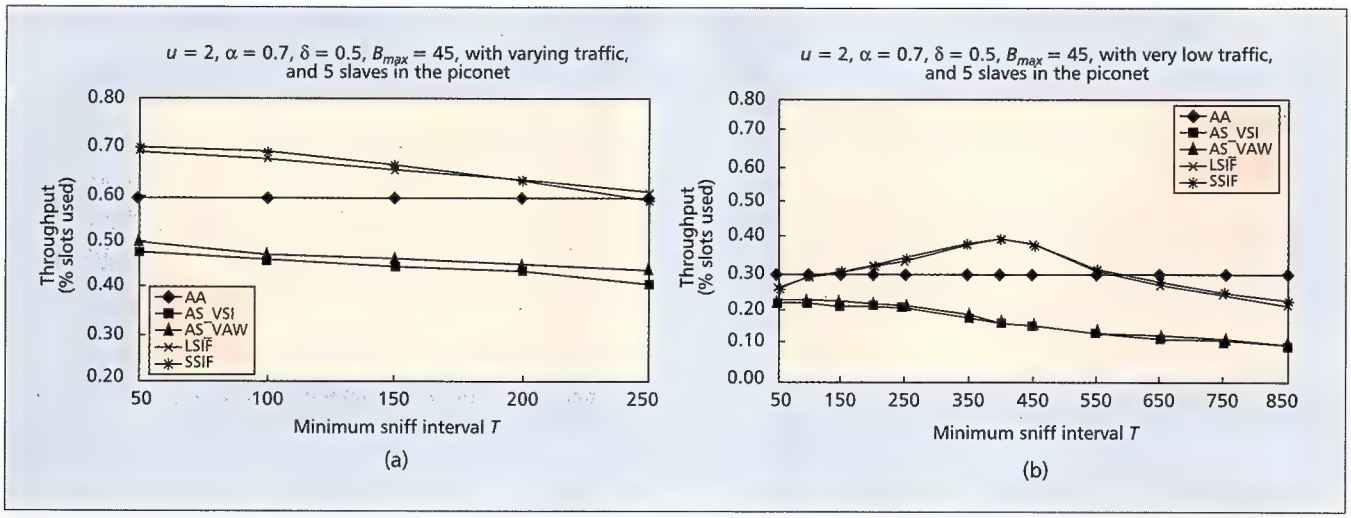
■ Figure 6. *The effect of a)* α *and b)* δ *on LSIF and SSIF under varying traffic load.*

traffic load. In Fig. 7b, a value of $u = 4$ will perform the best. This justifies that our 2D matrix representation of M is flexible enough to schedule sniffing slots for hosts with both high and very low traffic loads.

In Fig. 8a, we investigate the effect of T, which represents the smallest possible sniff interval, under the traffic patterns of TypeI(0.2–0.2), TypeI(0.19–0.01), TypeI(0.01–0.19), TypeII(0.19–0.01), and TypeII(0.19–0.01). It shows that the throughput of LSIF and SSIF will slightly decrease as T grows, so a T between 50 and 100 will be proper. The reason for the degradation in throughput is that we activate the evaluator based on the value of T. A larger T will trigger the protocol to reevaluate its traffic load less frequently. The inaccuracy in load estimation will cause slot waste. We believe this prob-



■ Figure 7. *The effect of* u *on LSIF and SSIF under a) varying traffic and b) very low traffic load.*



■ Figure 8. *The effect of* T *on LSIF and SSIF under a) varying traffic and b) very low traffic load.*

lem can be fixed by using a different rule to trigger our evaluators, which will be directed to future research.

Another experiment to investigate the impact of T is in Fig. 8b, where we try low traffic patterns: TypeI(0.2–0.2), TypeI(0.001–0.001), TypeI(0.002–0.002), TypeII(0.002–0.005), and TypeII(0.002–0.005). Similar to the earlier observations, lower traffic loads require larger $2^u \cdot T$. This is why we see continuous improvement before $T \leq 400$. Once T is too large, our evaluators will react to traffic changes too slowly, causing degradation in throughput. The results from Figs. 8a and 8b suggest T between 100 and 200.

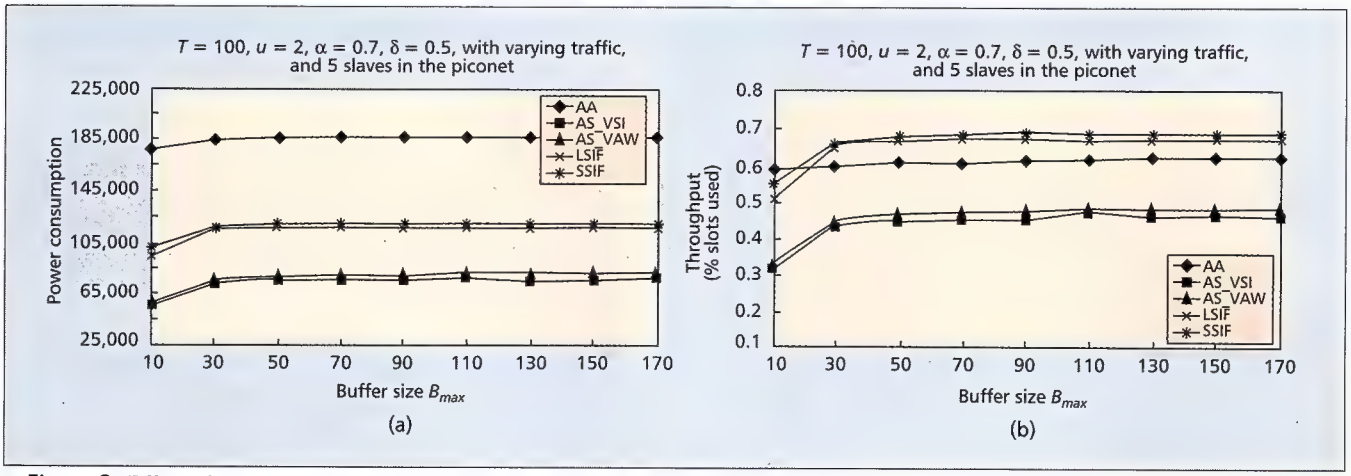
In Fig. 9, we set our parameters as recommended above and look at the impact of buffer spaces, B_{max}. We observe both power consumption and network throughput against different buffer spaces. It shows that the throughput climbs as B_{max} increases, up to $B_{max} = 50$. Once $B_{max} \geq 50$, the throughput remains almost the same. Thus, a buffer space between 30 and 50 is proper. With $B_{max} \approx 30$, LSIF and SSIF increase system throughput by around 16.7 percent compared to AA, while reducing power consumption by around 37.8 percent compared to AA.

Conclusions

We propose an adaptive and efficient protocol for managing the low-power sniff mode in Bluetooth. Two essential parts of our protocol are the evaluator and scheduler, which are responsible for measuring how well slaves utilize their sniff-attempt slots and arranging the sniff parameters for slaves in the sniff mode. A new representation based on a two-dimensional matrix is proposed to maintain slaves' sniff-attempt slots, which are conceptually infinite sequences of periodical slots. Two searching strategies, Longest Sniff Interval First and Shortest Sniff Interval First, are proposed to look for available sniff-attempt slots in the two-dimensional matrix. Our simulation results indicate that, with proper settings and buffer spaces, the protocol can potentially improve network throughput, while reducing power consumption, over a naive always active round-robin protocol.

Acknowledgments

Y.-C. Tseng would like to thank the Lee and MTI Center for Networking Research at NCTU, the Ministry of Education (contract nos. 89-H-FA07-1-4 and 89-E-FA04-1-4), and the National Science



■ Figure 9. *Effect of* B_{max} *under varying traffic load.*

Council, Taiwan (contract no. NSC90-2213-E009-154) for financially supporting this research.

References

[1] Bluetooth SIG http://www.bluetooth.com, "Bluetooth Specification v1.1," Feb. 2001.
[2] J. C. Haartsen, "The Bluetooth Radio System," *IEEE Pers. Commun.*, Feb. 2000.
[3] J. C. Haartsen and S. Mattisson, "Bluetooth — A New Low-Power Radio Interface Providing Short-Range Connectivity," *Proc. IEEE*, vol. 88, Oct. 2000.
[4] R. Bruno, M. Conti, and E. Gregori, "WLAN Technologies for Mobile Ad Hoc Networks," *IEEE Proc. 34th Hawaii Int'l. Conf. Sys. Sci.*, 2001.
[5] S.-L. Wu, Y.-C. Tseng, and J.-P. Sheu, "Intelligent Medium Access for Mobile Ad Hoc Networks with Busy Tones and Power Control," *IEEE JSAC*, vol. 18, Sept., 2000, pp. 1647–57.
[6] R. Ramanathan and R. Rosales-Hain, "Topology Control of Multihop Wireless Networks Using Transmit Power Adjustment," *IEEE INFOCOM*, 2000, pp. 404–13.
[7] C.-F. Huang *et al.*, "Increasing the Throughput of Multihop Packet Radio Networks with Power Adjustment," *Int'l. Conf. Comp. Commun. and Nets.*, 2001.
[8] J. H. Ryu, S. Song, and D.-H. Cho, "A Power-Saving Multicast Routing Scheme in 2-tir Hierarchical Mobile Ad-Hoc Networks," *IEEE VTC*, vol. 4, 2000, pp. 1974–78.
[9] S. Singh, M. Woo, and C. S. Raghavendra, "Power-Aware Routing in Mobile Ad Hoc Networks," *Int'l. Conf. Mobile Comp. and Net.*, 1998, pp. 181–90.
[10] H. Woesner *et al.*, "Power-Saving Mechanisms in Emerging Standards for Wireless LANs: The MAC Level Perspective," *IEEE Pers. Commun.*, June 1998, pp. 40–48.
[11] I. Chakrabory *et al.*, "MAC Scheduling Policies with Reduced Power Consumption and Bounded Packet Delays for Centrally Controlled TDD Wireless Networks," *IEEE ICC*, 2001.
[12] A. Das *et al.*, "Enhancing Performance of Asynchronous Data Traffic over the Bluetooth Wireless Adhoc Network," *IEEE INFOCOM*, 2001.
[13] S. Garg, M. Kalia, and R. Shorey, "MAC Scheduling Policies for Power Optimization in Bluetooth: A Master Driven TDD Wireless System," *IEEE VTC*, 2000.
[14] M. Kalia, D. Bansal, and R. Shorey, "Data Scheduling and SAR for Bluetooth MAC," *IEEE VTC*, 2000.

Biographies

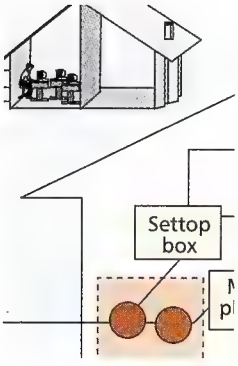
TING-YU LIN (tylin@csie.nctu.edu.tw) received her B.S. degree in computer science from National Chiao-Tung University, Taiwan, in 1996. She is currently a Ph.D. candidate at the Department of Computer Science and Information Engineering at the same university. Her research interests include wireless communication, mobile computing, personal-area networks, and energy conservation.

YU-CHEE TSENG (yctseng@csie.nctu.edu.tw) is currently a full professor in the Department of Computer Science and Information Engineering, National Chiao-Tung University, Taiwan. He has served as a Program Committee Member for several international conferences and as a Guest Editor for several journals, including *IEEE Transactions on Computers*, *Wireless Communications and Mobile Computing*, *Wireless Networks*, and *Journal of Internet Technology*. His research interests include wireless communication, network security, parallel and distributed computing, and computer architecture. He is a member of the IEEE Computer Society.

Our simulation results have indicated that, with proper settings and buffer spaces, the protocol can potentially improve network throughput, while reducing power consumption, compared to a naive always active round-robin protocol.

A POWER LINE COMMUNICATION NETWORK INFRASTRUCTURE FOR THE SMART HOME

YU-JU LIN, HANIPH A. LATCHMAN, AND MINKYU LEE, UNIVERSITY OF FLORIDA
SRINIVAS KATAR, INTELLON CORPORATION



Low voltage electrical wiring has largely been dismissed as too noisy and unpredictable to support high-speed communication signals. Recent advances in communication and modulation methodologies have spawned novel protocols capable of supporting power line communication networks at speeds comparable to wired LANs.

ABSTRACT

Low voltage electrical wiring in homes has largely been dismissed as too noisy and unpredictable to support high-speed communication signals. However, recent advances in communication and modulation methodologies as well as in adaptive digital signal processing and error detection and correction have spawned novel media access control and physical layer protocols, capable of supporting power line communication networks at speeds comparable to wired local area networks. In this article we motivate the use of power line LANs as a basic infrastructure for building integrated smart homes, wherein information appliances ranging from simple control or monitoring devices to multimedia entertainment systems are seamlessly interconnected by the very wires that provide them electricity. By simulation and actual measurements using "reference design" prototype commercial powerline products, we show that the HomePlug MAC and PHY layers can guarantee QoS for real-time communications, supporting delay-sensitive data streams for smart home applications.

INTRODUCTION

Many next-generation appliances are being equipped with processors featuring sophisticated communication capabilities. For instance, on April 7, 2001, IBM and Carrier announced plans to produce an air conditioner with Java support that can email manufacturers regarding errors, and will allow users to remotely send commands to the unit to adjust temperatures or shut it down. *Smart homes* will eventually have many types of information appliances (IAs) communicating among themselves and with the outside world. Soon, many of these IA devices are expected to have multimedia capability. Supporting multimedia communication for these IA devices will be of crucial importance for the intelligent homes of the future.

Providing the right infrastructure for connecting these IA devices will be a major need. For home applications, this infrastructure must be easy to set up, inexpensive to install and maintain, and must perform well. Homeowners are generally not network experts, and a typical high-performance network is too complicated for casual daily usage. The supporting infrastructure should be as easy as possible to set up, and the effort to maintain this infrastructure should be minimal. Many existing networking technologies compete to support this mission. For example, a comprehensive Ethernet network can be constructed by installing UTP-5 special cabling around the house. Alternatively, wireless networks such as *802.11x*, *Bluetooth*, and *HomeRF* can be constructed by installing multiple interconnected wireless access points(WAPs) and base stations within the home. However, the IA devices themselves would need wireless capabilities, and the above three infrastructures all require significant effort and cost to build up the networks externally. Phone line networks such as *HomePNA* [1] may seem attractive, but the convenience of mobility is limited by available phone sockets in a home. An extensive study of other infrastructure options and technologies appropriate for a home network is given in [2].

In this article we advocate the direct use of the existing electrical wiring and outlets as the medium for data communication within the home. Using power lines as the network infrastructure has many advantages over other technologies. First, no new wires are needed since the IA devices will communicate over the very wires that provide them electrical power. Second, there are many access points (power sockets) in a home (four or more per room). Currently, *power line communication* (PLC) as specified by the *HomePlug* 1.0 standard [3] provides a 14 Mb/s raw data rate, which is adequate for daily IA device communication. It also has a built-in QoS protocol, making it attractive for real-time streaming applications. Future generations of the PLC protocol will provide 100 Mb/s PLC services to support high-quality digital multimedia. Finally, the cost to build a power line network is low compared to that of other technologies. For example, it was observed that the 802.11b wireless network card has approximately the same street price as the HomePlug network card (about $70). It is expected that with mass production requiring no expensive RF components, the cost of the PLC cards will be about 50 percent that of comparable wireless cards. Moreover, the cost of a required 802.11b base station

1070-9916/02/$17.00 © 2002 IEEE

is high (more than $150). *100BaseT Ethernet* has the highest performance/cost ratio, but requires new cables and expensive installation. Table 1 shows costs and other various characteristics of home network technologies. Installation costs for 10/100BaseT, which are high, are not shown.

From a marketing perspective, the less expensive and easier-to-use PLC home network is becoming more attractive, and the potential market is huge. The Yankee Group estimates that at least 21 million households in the United States are interested in home networking and that 12.4 million would like to implement in-home networks within the next year. According to Parks Associates, 30 million households in the United States will have fast Internet connections by 2004, and 17 million of them plan to have home networks.

In the past, power lines were considered unacceptable for signal transmission, since the channel is subject to a lot of noise, interference, and fading. However, the appeal of using the existing power line as a transmission medium for data exchange was too great to be ignored. The advancement of signal modulation technologies, digital signal processing, and error control coding [5] has minimized the restrictions of channel imperfections, and high-speed signal transmission through power lines is now feasible [3].

Using the existing power line infrastructure as the medium for supporting IA communication requires careful design of the overlaid communication systems in order to provide acceptable communication services. It is desired, for example, that when watching digital TV while downloading data from the Web, there will be no delay jitter in the video quality. Current research shows the maximum raw data rate of first-generation PLC is about 14 Mb/s [3]. However, the effective data rate is expected to be around 10 Mb/s after compensation for impairments and error corrections. On the other hand, research is currently underway to develop PLC chips that operate at 100 Mb/s with average throughput in the range of 30–60 Mb/s.

In this article we investigate the performance of multimedia over power line networks using simulation studies and actual measurements on a HomePlug 1.0 compliant PLC network. We are particularly interested in measuring the PLC network raw data rate, TCP performance, and the performance impact when quality of service (QoS) support is involved. We are also interested in analyzing the network performance with different traffic types, including continuous media data streams (i.e., soft real-time traffic).

We first built a network simulator that generates various types of traffic, and then applied the same scenarios to a real-world PLC network and a simulation model. The performance comparison between the simulation results and real-world performance is given in this article. A maximum throughput of 8.08 Mb/s for UDP was obtained from our simulation, while a 6.21 Mb/s TCP throughput was observed in the real-world PLC network experiment. The results show that PLC networks can successfully deliver real-time traffic concurrently with traditional data traffic.

Our contributions in this article also include modeling human behavior in the use of IA devices, modeling the types of traffic generated by IA devices communicating over the power line, and measuring the performance of real applications over PLC networks. We also give a description of a practical implementation of the PHY and MAC layers for PLC networks as well as associated simulation results.

Technology	Media	Data rate	QoS Support	Cost
10BaseT	UTP	10 Mb/s	No	$10
100BaseT	UTP	100 Mb/s	No	$10
Bluetooth	Wireless	1 Mb/s	Yes	$5/chip
HomeRF 2.0	Wireless	10 Mb/s	Yes	$70
802.11b	Wireless	11 Mb/s	No	$80
HomePNA 2.0	Phone line	10 Mb/s	No	$60
HomePlug	Power line	14 Mb/s	Yes	$70

■ **Table 1.** *Technology comparison.*

In what follows we discuss the background for implementing PLC networks, typical applications of PLC networks, and the human behavioral model using these appliances. We present the physical limitations of power line channels, and describe a practical signal modulation scheme and MAC protocol as used in the ***HomePlug*** 1.0 protocol proposed by the HomePlug Powerline Alliance. We compare the simulation results with real-world PLC network performance. Finally, we present some conclusions and suggestions for future work.

Applications over Power Line Communication

Power Line Communication Application in a Home

Traditionally, power lines are used for conveying electrical power to devices. Power lines were not designed for delivering high-frequency signals, so the electrical and frequency response requirements of a power line are not as critical as those of data network cabling. The poor quality of a power line is not ideal for signal transmission because the channel contains noise and interference. The medium is made of different conductor types; therefore, a variety of characteristic impedances will be encountered. Furthermore, the network terminal impedance will tend to vary with frequency and time as the consumer's load pattern and load types vary. Impedance mismatch causes a multipath effect, resulting in deep notches at certain frequencies. These channel imperfections make signal modulation over a power line difficult [4]. However, the advancement of signal modulation and error control coding techniques now make power line communication possible.

The common power line topology of a North American home is depicted in Fig. 1a. The figure shows a tree-like power line topology in a house. Typically, there are two power line trunks, one 110 V and the other 220 V. Each power line trunk can be divided into several branches. PLC

aims to transmit data packets over these branches and trunks. The topology of the power line network and the convenience of its power sockets as potential access points make it a good candidate for smart home IA device networking.

Table 2 shows the results of a survey from which we inferred usage and traffic patterns generated by typical IAs. The table also suggests some current and future PLC applications. For instance, when merchandize is advertised on a digital TV service, the product information (e.g., the barcode or Webpage) can be downloaded to your computer through a power line. Afterward, you can send your order information from the computer to the supplier or use the downloaded URL to browse the product Web page and get more details. We also anticipate the ability to record music or videos through a power line. For example, when a song is being broadcast on TV or a music channel, you can download the song directly to an MP3 player through the power line. Another application is the opportunity to record digital video directly into a PC or even a digital VCR.

Other applications of IAs can easily be accomplished using a PLC network. For example, a refrigerator can order food through the power line network according to its inventory, or send cooking instructions to the microwave. A smart oven can send predicted environmental temperature information to the air conditioner through a power line, allowing the air conditioner to pre-adjust the temperature and keep rooms comfortable.

The applications over PLC are not only for novel IA devices. PLC as a home network facilitates data exchange between traditional data processing devices such as PCs and computer peripherals. IA devices that talk with PCs are also possible. For example, sending multimedia data from TVs or VCRs to PCs can easily be done by a PLC network, but is difficult with other infrastructure technologies. Home security can also be implemented by PLC so that a digital camera installed on the front door can send video to the TV.

Table 2 also gives an estimate of the daily traffic volume generated by typical IA applications. These values are based on likely information size. For example, the instruction size the refrigerator sends to the microwave in row 1 is estimated by the number of steps required to cook the food (1 byte), the cooking time for each step (4 bytes for each step), the power level for each step (2 bytes), and the packet header size. Added together, the entire instruction size is 160 bytes. Row 7 exemplifies storing digital music from a computer to an MP3 player. The 50 Mbyte traffic volume is calculated from the number of songs in an album, the length of a song (5 min), the encoded data rate (128 kb/s), and the packet header size. The frequency and time period during which each event occurs are also shown. By using this data and typical household dynamics for concurrent events, we can generate a traffic flow for the power line network for a typical day.

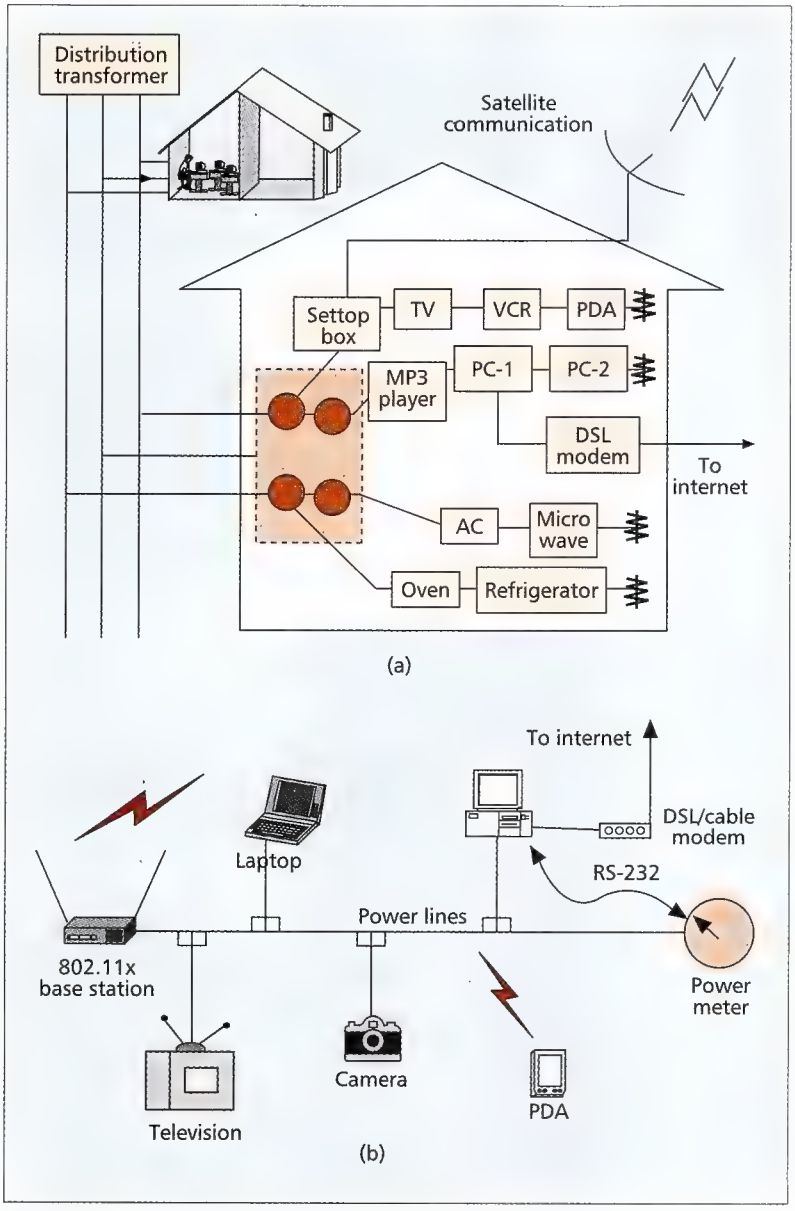


Figure 1. *a) Power line topology in a North American home; b) An example of using one of the computers as the PLC, DSL or cable modem router.*

Internet Bridging

Currently, in-home PLC networks have to rely on other technologies to send data to the Internet and communicate with mobile devices. Most of the homes in the United States will eventually be equipped with broadband connections like DSL or cable modem services. To share the broadband Internet connection with PLC-capable devices, we can add a PLC Internet router to the PLC network. One possible setup is depicted in Fig. 1b.

In this figure, a desktop computer acts like a data center. Devices that need to communicate with other devices on the Internet will send data to the desktop PC via the power line. The desktop PC decides whether to send it to the Internet. In the future, an IA routing device may be unnecessary. Researchers are developing a solution to make PLC home networks talk directly with other homes, power plants, and the Internet using the external distribution power line. Such a network infrastructure for Internet access would be especially attractive to developing countries,

Row	From node	To node	Estimated data size	Frequency	Possible time period
1	Refrigerator	Microwave	160 bytes	2 times a period	7:00–9:00, 11:00–1:00, 17:00–19:00, 21:00–23:00
2	Microwave	AC	72 bytes	2 times a day	7:00–9:00, 11:00–1:00, 17:00–19:00, 21:00–23:00
3	TV	Refrigerator	750 bytes	3 times a day	11:00–1:00, 17:00–23:00
4	TV	VCR	11 kbytes	3 times a day	11:00–1:00, 17:00–23:00
5	TV	Computer	360 bytes	3 times a day	11:00–1:00, 17:00–23:00
6	TV or Settop box	PDA or MP3 player	15 Mbytes		11:00–1:00, 17:00–23:00
7	Computer	PDA or MP3 player	50 Mbytes	1 time a day	11:00–1:00, 17:00–23:00
8	Computer	Computer	60–180 Mbytes	1 time a day	6:00–24:00
9	Settop box	Computer	320–640 Mbytes	1 time a day	11:00–1:00, 17:00–23:00
10	Computer	Internet	44–131 Mbytes	1 time a day	11:00–1:00, 17:00–23:00
11	VCR	Computer	320–640 Mbytes	1 time a day	6:00–24:00
12	Front door camera	Computer	110–1100 Mbytes	3 times a day	6:00–24:00

■ **Table 2.** *The amount of application traffic in a home.*

since no additional expenditure is needed for data network infrastructures.

To support data exchange with mobile devices, PLC networks will also need to cooperate with wireless networks. The easiest way to achieve this is to make the 802.11x base station PLC compatible. The base station is treated as an ordinary IA device with a PLC chip built in (Fig. 1b). Mobile devices with wireless capability can then talk to devices attached to the power line. This is especially ideal when communication is desirable but large coverage areas require multiple interconnected wireless access points; the interconnection is then provided with "no new wires" using the existing power line infrastructure which would be needed to power the WAPs in any event.

The above PLC applications require a properly designed protocol. In addition, to make the PLC network real-time traffic friendly, special care is needed to support delay-sensitive traffic. In the following sections, we discuss the physical limitations of power line channels and then describe a robust power line protocol.

PLC Design Issues

Physical Limitations

A power line is used for transmitting 50 or 60 Hz signals but was not designed to convey high frequency signals such as the 20 MHz communication signal used in the HomePlug 1.0 protocol. A power line channel is somewhat like a wireless channel: both suffer from noise, fading, multipath, and interference. Power line noise is produced by the operation of electrical devices. Fading, multipath, and interference are caused by the imperfection of power line channels.

Typical attenuation characteristics in power line channels are given in [6]. The author reports that even when all devices are unplugged, the noise still persists, and this drastic variation of attenuation is hostile to PLC. Furthermore, the Federal Communications Commission (FCC) also limits the available bandwidth for communication purposes. In compliance, the usable bandwidth in the HomePlug standard is 25 MHz. An extensive study of power line channel characteristics and design issues is given in [7].

To conquer these problems, robust signal modulation and data coding are needed.

Signal Modulation

To modulate digital signals onto the power lines, we can use many of the same techniques widely implemented in wireless communication. Basic modulation techniques such as phase shift keying (PSK), frequency shift keying (FSK), minimum shift keying (MSK), and Gaussian minimum shift keying (GMSK) can be used for low-data-rate communication. Other more advanced techniques such as M-ary PSK (MPSK), M-ary quadrature amplitude modulation (MQAM), M-ary FSK (MFSK), and orthogonal frequency-division multiplexing (OFDM) can be used when higher data rates are desired. A thorough study of signal modulation over power lines is given in [8].

OFDM was adopted by the HomePlug Powerline Alliance because of its robustness to noise and the fact that it is a parallel data transmission method using a number of parallel FDM subbands. The main problem in using OFDM on wireless networks is frequency offset, caused by the Doppler effect when the user is moving. The Doppler effect will cause performance degradation, but in a power line network there are no moving devices, and thus no Doppler effect. The other problem is timing offset, which can be mitigated by offset estimation and compensation.

Spread spectrum signal modulation is different. Since the useful bandwidth in the power line channel is under 25 Mhz, the effect of spread spectrum modulation is considered limited. Using single carrier modulation on the power line is possible, but equalizers could be needed

This privacy protection seems adequate but stronger encryption may be needed when power line networks are adopted for office environments or apartment building and hotels. We believe that stronger privacy protection should be implemented in the physical layer, so that hackers cannot easily break the code.

to reduce the delay spread effect, and the associated cost is high.

In order to cope with the wide variation in channel conditions, the physical layer protocol (PHY) for PLC must be adaptive, intelligently using more robust modulation and coding schemes, with lower data rates as needed. In addition, critical protocol management information requires high fidelity forward error correction (FEC) coding to ensure that the protocol functions correctly in the worst case situations.

MAC Layer Protocols

In PLC home networks, the power line media can be accessed by multiple devices simultaneously. To decide which device gets the floor to send its data, a medium access control (MAC) protocol is needed. There are many existing protocols that can be implemented on the power line network. Carrier sense multiple access, CSMA with collision avoidance (CSMA/CA), time-division multiple access (TDMA), and hybrid protocols such as TDMA+CSMA are all potential candidates.

The most popular wired MAC protocol, CSMA with collision detection (CSMA/CD), could also be applied on a power line network. However, the large variation in noise on the power line makes collision detection very difficult. This characteristic is again very similar to a wireless network, so some have applied the CSMA/CA protocol as suggested in IEEE802.11 to the power line network. However, the hidden node problem arises when the signal travels through different power lines with highly variable attenuation. To address this problem an RTS/CTS scheme could be implemented. Although the RTS/CTS scheme mitigates the hidden node problem, it degrades network performance.

The benefit of using TDMA is that it provides an upper bound on access delay; thus, QoS is guaranteed. However, the difficulty in generating a synchronized clock signal in power line networks between devices remains a problem. Other hybrid protocols like TDMA+CSMA provide QoS capabilities, but network efficiency and beacon generation between TDMA slots and CSMA/CA slots remain unsolved. A detailed discussion of the hybrid TDMA+CSMA/CA protocol is provided in [10]. HomePlug 1.0 also provides some level of QoS support in the multiple priority levels that can be used in conjunction with virtual LAN (VLAN) tagging.

The issue of privacy of power line networks is important to their practicality. Like wireless channels, power line network channels should be treated as open, and, as with all open channels, nothing prevents a device from receiving signals. To provide a secure network environment, the HomePlug Powerline Alliance defined a 56-bit DES encryption mechanism. Once a signal is encrypted, a device with a different encryption key cannot interpret it, and privacy is achieved.

This privacy protection seems adequate, but stronger encryption may be needed when power line networks are adopted for office environments or apartment buildings and hotels. We believe stronger privacy protection should be implemented in the physical layer so that hackers cannot easily break the code.

Performance Results and Analysis

In this section we report the measurements observed using an event-based C program to simulate a HomePlug 1.0 power line network. All scenarios assume QPSK and a 3/4 coding rate on various links, and a maximum TCP segment size of 1460 bytes. In this simulation, we use UDP, TCP, and voice over IP (VoIP) traffic. UDP traffic is generated with an exponential interarrival time with a 100 μs average. The UDP packet size is assumed to be a constant 1460 bytes with low priority. TCP traffic is also generated with exponential interarrival time with 100 μs average, and we assume that TCP traffic sources always have data to send. TCP traffic is treated as low priority packets. Every time a node has a chance to send, it is allowed to send the maximum segment size of 1460 bytes without headers. VoIP is isochronous traffic with a 20 ms interval. The packet size of VoIP is 160 bytes and is assigned the high priority.

Simulation Results

In Table 3, we provide the simulation results of a power line network. UDP traffic simulation scenario 1 shows the best throughput in our simulations since there is no contention at all. Table 3 also shows that channel contention with two and three UDP nodes causes a modest reduction in channel throughput.

In the TCP traffic simulation, although scenario 1 has only one traffic source, the bandwidth must be shared with data and response frames (e.g., ACK packets); thus, it provides lower performance than the UDP traffic simulation. The MAC throughput represents the total number of transmitted bytes divided by the simulation time regardless of successful delivery. The TCP throughput includes only the successfully delivered data and ACKs.

The third metric we provide in Table 3 is the PLC simulation results of one VoIP and multiple UDP connections. The high-priority VoIP always wins the contention and the UDP nodes can send packets only when there is no VoIP traffic. In this simulation, the queuing delay refers to the time a packet waits in a queue before it enters the transmit buffers. The net delay is the total time for which a packet propagates in the networks. Only low priority packets are considered for this delay because the high priority packets will be delivered as soon as they appear in the queue.

Table 3 also shows the simulation results of one VoIP and multiple TCP connections. The throughput of VoIP is only 80 kb/s, and hence the total throughput is dominated by the TCP component.

Real-World PLC Network Performance

In addition to simulating the performance of the HomePlug Powerline Alliance protocol, we were also able to construct a real PLC network using "reference designs" of actual commercial HomePlug devices. Since there are currently no real IA devices with PLC capability, we used traditional network applications (i.e., ftp and streaming multimedia content) as the basis for measuring PLC network performance.

Throughput of multiple UDP traffic			
	Scenario 1 (1 UDP)	Scenario 2 (2 UDP)	Scenario 3 (3 UDP)
MAC throughput	8.08 Mb/s	7.46 Mb/s	7.46 Mb/s
Throughput of multiple TCP traffic			
	Scenario 3 (3 TCP)	Scenario 2 (2 TCP)	Scenario 1 (1 TCP)
MAC throughput	6.16 Mb/s	6.15 Mb/s	6.12 Mb/s
TCP throughput	5.92 Mb/s	5.91 Mb/s	5.88 Mb/s
Throughput of one VoIP and multiple UDP traffic			
	Scenario 1 (VoIP + 1 UDP)	Scenario 2 (VoIP + 2 UDP)	Scenario 3 (VoIP + 3 UDP)
MAC throughput	7.89 Mb/s	7.33 Mb/s	7.29 Mb/s
Queuing delay	0.25 ms	0.25 ms	0.25 ms
Net delay	3.00 ms	3.00 ms	2.75 ms
Throughput of one VoIP and multiple TCP traffic			
	Scenario 1 (VoIP + 1 TCP)	Scenario 2 (VoIP + 2 TCP)	Scenario 3 (VoIP + 3 TCP)
MAC throughput	6.04 Mb/s	5.85 Mb/s	5.77 Mb/s
TCP throughput	5.72 Mb/s	5.54 Mb/s	5.45 Mb/s
Queuing delay	0.25 ms	0.25 ms	0.25 ms
Net delay	3.25 ms	3.25 ms	3.25 ms

Table 3. *Power line network simulation results.*

The PLC network successfully delivered both low and medium bit rate streaming videos. We did not observe any packet drops during the experiments. The results met our expectations, since the peak data rate was only 4185 kb/s.

In this experiment, there were four desktop computers. A 450 MHz Pentium II desktop computer (PC-2 as a file server) is equipped with 128 Mbytes RAM, a 3COM fast Ethernet card, and a PLC PCI card. Two 700 MHz Pentium III desktop computers (PC-3 and PC-4) are both equipped with 256 Mbytes RAM, and PLC PCI cards. A 266 MHz Pentium MMX desktop computer(PC-1) is equipped with 64 Mbytes RAM, and a 3COM fast Ethernet card.

The PC-1 computer is connected to an Ethernet-to-power-line bridge, which converts packets generated from the Ethernet card into PLC compatible packets, and vice versa. All computers are connected to power lines.

In this experiment, we seek to determine the performance of the PLC network in handling streaming video and large file transfers.

Performance for Delivering Streaming Video — We first examined the ability of the PLC network to deliver real-time traffic. Four video files are involved in this experiment. The first file is encoded in Real media format with a bit rate of 550 kb/s; the second is encoded with a bit rate of 1396 kb/s, the third is encoded at 2 Mb/s, and the fourth is an MPEG2 video file with variable bit rate, and the average bit rate is 8 Mb/s. In the first experiment, three client computers simultaneously issued requests for low-bit-rate (550 kb/s) video service to the file server. In the second experiment, the same procedure was executed, but a medium bit rate (1394 kb/s) video service was requested. In the third experiment, the three clients requested a high-bit-rate (2 Mb/s) video service. Finally, the MPEG2 video service request was issued by the PC-3. The experimental results are shown in Table 4.

The PLC network successfully delivered both low- and medium-bit-rate streaming videos. We did not observe any packet drops during the experiments. The results met our expectations, since the peak data rate was only 4185 kb/s. We did another experiment to further investigate the performance of PLC network in delivering streaming video. A 2 Mb/s MPEG-1 file is used in this experiment. As the video begins, a significant video freeze-then-go (halting) phenomenon was observed, causing staccato playback. After several seconds(3–5 s) the freeze-then-go phenomenon disappeared.

In the case of MPEG2 video file, the average data rate is 8 Mb/s. During the experiment, a significant "video staccato" phenomenon was observed. To exclude the possibility that the observed phenomenon was caused by the client computer's hardware capability, the experiment was repeated with the same configuration, while connected to a fast Ethernet. During that experiment, no such phenomenon (halting playback) was observed.

Performance of Elastic Data Traffic — The occurrence of the momentary video freezing phenomenon during playback of variable bit rate streaming is likely because the aggregated data rate was close to or exceeded the PLC network capacity. To understand the real throughput of a PLC network, we conducted another experiment. A 215,502,106 byte file was placed on the server running an FTP daemon. (The file size was chosen to minimize hardware uncer-

Aggregated traffic volume decreased as the number of connections increased from one to two. This phenomenon is because of the ACK packets and the packet overhead increase as the number of connections increased. Although the network utilization improves, the improvement cannot compensate for the loss due to these overheads.

Performance of delay-sensitive traffic

	Number of connections	Aggregated bit rate	Packet drop	Delay jitter
Low bit rate	1	550 kb/s	No	No
	2	1100 kb/s	No	No
	3	1650 kb/s	No	No
Medium bit rate	1	1395 kb/s	No	No
	2	2790 kb/s	No	No
	3	4185 kb/s	No	No
High bit rate	1	2000 kb/s	No	No
	2	4000 kb/s	No	No
	3	6000 kb/s	N/A	Moderate
	Number of connections	**Average bit rAte**	**Environment**	
Variable bit rate	1	8 Mb/s	PLC network	
	1	8 Mb/s	Fast Ethernet	

Performance of elastic data traffic

	Number of connections	Average bit rate
Elastic data traffic	1	6.21 Mb/s rate
	2	6.15 Mb/s rate
	3	6.27 Mb/s rate

Performance of combined delay sensitive traffic and elastic data traffic

	Connections	Aggregated bit rate
Hybrid data traffic	One FTP connection and one video service	6.26 Mb/s rate
	Two FTP connections and one video service	5.92 Mb/s rate

■ **Table 4.** *Real-world PLC network performance.*

tainty and human error.) Client computers made FTP requests for the file. We tested different numbers of FTP connections, up to three, using individual client machines in our PLC network. The experimental results are also given in Table 4.

The aggregated traffic in the table is calculated by adding all observed data rates of all connections. The experimental results show that the real PLC network performance is about 6 Mb/s. When there is only one FTP connection, the observed throughput is 6.21 Mb/s. By our analysis, one TCP connection will not fully utilize the PLC network, because the server has to stop if no ACK packets are received from the client.

Aggregated traffic volume decreased as the number of connections increased from one to two. This phenomenon is because of the ACK packets and the packet overhead increase as the number of connections increased. Although the network utilization improves, the improvement cannot compensate for the loss due to these overheads.

When we increased the number of connections from two to three, the PLC network had the highest throughput of 6.27 Mb/s. This is because the network utilization increased as the number of connections increased, which compensated for the packet overhead and ACK overhead.

These experimental results explain the phenomenon of momentary DVD video freezing playback. The requested bandwidth for DVD streaming exceeded the maximum bandwidth the present PLC network can provide.

Performance of Combined Delay-Sensitive Traffic and Elastic Data Traffic — Although we could not explore the QoS service and packet priority provided by the real PLC network, we were eager to learn the effect of mixed traffic on the PLC network. This experiment was conducted as follows: The file server provided two services: one for streaming video with bit rate 550 kb/s, and the other one for file transfer with a file size of 215,502,106 bytes. PC-1 requested streaming video, while PC-2 and PC-3 requested the file transfer. Each experiment lasted 285 s (i.e., the length of the video file), after which both video player and FTP client were forced to stop. Table 4 shows our experimental results.

When the number of FTP connections increased, the observed data rate decreased, as was the case in the previous experiment. However, the overall average data rate was comparable to the case of multiple FTP traffic.

Conclusion

The emergence of information appliances for the smart homes of the future will undoubtedly make our lives much more comfortable than ever. However, the infrastructure that supports multimedia traffic and conventional elastic data traffic for communication among IA devices is a critical component of a smart home.

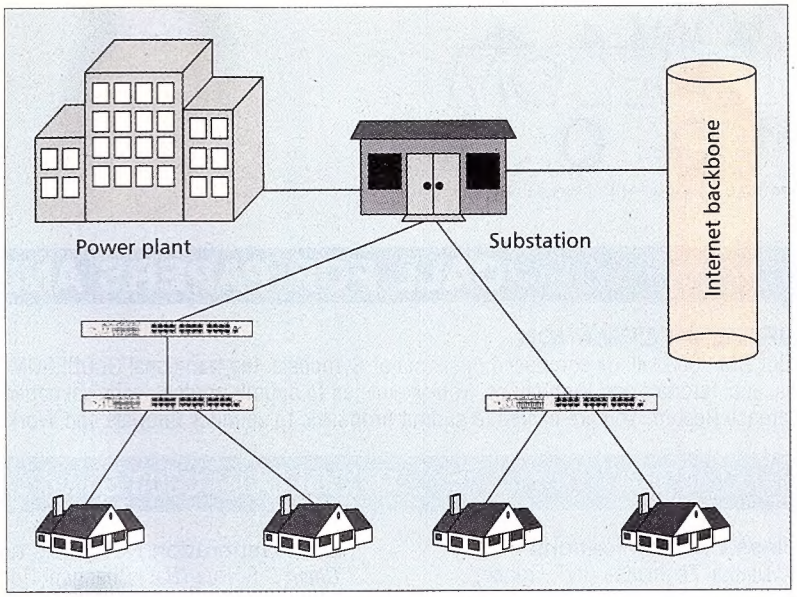
We advocate power line as the infrastructure for smart homes based on the convenience of the power sockets and the layout of the power line network existing in every home. At present, 6 Mb/s of bandwidth was measured through real-

world PLC network experiments. Our studies showed that the PLC network can provide three low-bit-rate or three medium-bit-rate multimedia streams concurrently with no packet drops and jitters. It also successfully delivered one low-bit-rate multimedia data stream and two large FTP file transfers concurrently with no packet drops and jitters.

In this article we discuss only PLC networks for communication *within* the smart home, but the ultimate goal could be the ability to connect to the Internet without dialing up to an Internet service provider's server, using electrical wiring only. This is illustrated in Fig. 2.

Private home networks are connected to substations, in which a powerline modem terminal service (PMTS) connects PLC networks within homes to the Internet backbone. The PLC network gateway for a private home network could be installed in the fuse box of that home and then connected to one or more repeaters. Repeaters are for increasing signal strength when the signal levels fall below some value.

We expect to see higher data rates in power line networks in the future as signal modulation technologies improve; however, issues like network security and network characteristics with a large number of nodes need further development. Further research on these issues is of critical importance when power line networks are applied to offices and large multi-user buildings.



■ **Figure 2.** *Connecting PLC networks to the Internet.*

References

[1] E. H. Frank and J. Holloway, "Connecting the Home with a Phone Line Network Chip Set," *IEEE Micro*, 39, vol. 20, no. 2, Mar.–Apr. 2000, pp. 27–37.

[2] S. Hughes and D. J. Thorne, "Broadband In-Home Networks," *BT Tech. J.*, vol. 16, no. 4, Oct. 1998, pp. 71–79.

[3] Intellon, http://www.intellon.com, on Nov. 20, 2001.

[4] J. S. Brown, "Physical Multipath Model for Power Distribution Network Propagation," *Proc. Int'l. Symp. Powerline Commun. and Its Apps.*, 1998, pp. 76–89.

[5] M. K. Simon and M.-S. Alouini, "A Unified Approach to the Probability of Error for Noncoherent and Differentially Coherent Modulations over Generalized Fading Channels," *IEEE Trans. Commun.*, vol. 46, no. 12, 1998, pp. 1625–38.

[6] C. K. Lim *et al.*, "Development of A Test Bed for High-Speed Power Line Communications," *Int'l. Conf. Power Sys. Tech.*, 2000, vol. 1, 2000, pp. 451–56.

[7] L. Weilin *et al.*, "Nature of Power Line Medium and Design Aspects for Broadband PLC System," *2000 Proc. Int'l. Zurich Sem. Broadband Commun.*, 2000, pp. 185–89.

[8] M. Karl and K. Dostert, "Selection of An Optimal Modulation Scheme for Digital Communications over Low Voltage Power Lines," *IEEE 4th Int'l. Symp. Spread Spectrum Tech. and Apps.*, vol. 3, 1996, pp. 1087–91.

[9] M. Sliskovic and B. Jeren, "Clock Frequency Synchronisation in OFDM System for Power Line Communications," *Proc. 1st Int'l. Wksp. Image and Sig. Processing and Analysis*, 2000, pp. 241–46.

[10] C. Romans and J. Tourrilhes, "A Medium Access Protocol for Wireless LANs Which Supports Isochronous and Asynchronous Traffic," *9th IEEE Int'l. Symp. Pers., Indoor and Mobile Radio Commun.*, vol. 1, 1998, pp. 147–52.

Biographies

YU-JU LIN received a Bachelor of Engineering degree from National Central University, Taiwan, in 1990, and an M.S. degree in computer and information engineering from Chung-Yuan Christian University, Taiwan, in 1995. He is currently a Ph.D. student in the Department of Electrical and Computer Engineering, University of Florida. His current research interests include multimedia communication and computing, power line communication, and high-speed networks.

HANIPH A. LATCHMAN [SM] (latchman@ufl.edu) was the 1983 Jamaica Rhodes Scholar and received his Ph.D. from Oxford University in 1986 and his B.S. degree (First Class Honors) from the University of the West Indies-Trinidad and Tobago in 1981. He teaches graduate and undergraduate courses and conducts research in the areas of control systems, and communications and computer networks, and has received numerous teaching and research awards, including the University of Florida Teacher of the Year Award and the IEEE 2000 Undergraduate Teaching Award with a citation "for innovative and inspirational teaching and advancing the use of information technology in education." He has published over 80 technical articles in the areas of his research. He is author of the books *Computer Communication Networks and the Internet* (McGraw Hill) and *Linear Control Systems — A First Course* (Wiley). He is also an Associate Editor for *IEEE Transactions on Education*.

SRINIVAS KATAR received a B.Tech. degree in electrical engineer in 1998 from the Indian Institute of Technology, Kanpur. From 1998 to 2000 he studied at the Department of Electrical and Computer Engineering, University of Florida, where he received an M.S. degree. In May 2000 he joined the Research and Development team at Intellon Corporation, Ocala, Florida. His current research interests include networking protocols, multicarrier communications, and error control coding.

MINKYU LEE received an M.S. degree in electrical and computer engineering from the University of Florida in 1999. He is currently working toward a Ph.D. degree in electrical and computer engineering at the University of Florida. His research interests are in the areas of power line home networks to support QoS and designing next-generation PLC protocols.

IEEE GLOBECOM 2003
COMMUNICATIONS: THE GLOBAL BRIDGE

IEEE GLOBECOM 2003

1-5 December
San Francisco, California
www.globecom2003.com

CALL FOR PAPERS AND TUTORIALS

GENERAL INFORMATION

GLOBECOM 2003 will be composed of six major Symposia, the traditional GLOBECOM General Conference addressing a set of current communication topics, and Tutorials and Workshops. You are invited to submit original technical papers for presentation at GLOBECOM 2003 and publication in the Conference Record. You are invited to submit proposals to conduct Tutorials and Workshops. Visit the website for further details.

GLOBECOM 2003 SYMPOSIA TITLES

Wireless Communications

Chair: Djamal Zeghlache (INT, France)
E-mail: djamal.zeghlache@int-evry.fr
Vice-chair: Josef Hausner (Infineon, Germany)
E-mail: Josef.hausner@infineon.com
Tracks: Ad-Hoc Wireless Networks, Wireless Multimedia, Satellite, Radio Technology, Wireless LAN's and Broadband Wireless.

Optical Networking and Systems

Chair: Ibrahim Habib (CUNY, USA)
E-mail: habib@ccny.cuny.edu
Vice-chair: Anthony Boucouvala (Bournemouth Univ., UK)
E-mail: tboucouv@bournemouth.ac.uk

Next Generation Networks and Internet

Chair: G.S. Kuo (Nat'l Chengchi University, Taiwan)
E-mail: gskuo@ieee.org
Vice-chair: Sherman Shen (Univ. of Waterloo, Canada)
E-mail: xshen@bbcr.uwaterloo.ca

Communication Theory

Chair: Alex M. Haimovich (NJIT, USA)
E-mail: haimovic@njit.edu
Vice-chair: Helmut Bölcskei (ETH, Swiss)
E-mail: boelcskei@nari.ee.ethz.ch

Signal Processing for Communications

Chair: Tomohiko Taniguchi (Fujitsu, Japan)
E-mail: tani@flab.fujitsu.co.jp
Vice-chair: Ron Smith (TRW, USA)
E-mail: Ron.P.Smith@trw.com

Communications Security

Chair: Victor K. Wei, CUHK, Hong Kong
E-mail: kwwei@ie.cuhk.edu.hk
Vice-chair: Yile Guo (Nokia, USA)
E-mail: Yile.Guo@nokia.com

GLOBECOM 2003 GENERAL CONFERENCE TOPICS

Chair: Mohsen Guizani (UWF, USA)
E-mail: mguizani@cs.uwf.edu

Vice-chair : Gang Wu (NTT DoCoMo, USA)
E-mail: wu@docomolabs-usa.com

- Cable-Based Delivery & Access Systems
- Communication Switching and Routing
- Communications Quality & Reliability
- Communications Security and Authentication
- Communications Software
- Communications Systems Integration & Modeling
- Computer Communications
- Distributed Computing
- Enterprise Networking
- Gigabit Networking
- Information Infrastructure
- Interconnections in High-Speed Digital Systems
- Multimedia Communications
- Network Operations & Management
- Quality of Service
- Modeling and Simulation
- Satellite & Space Communications
- Signal Processing & Communications Electronics
- Signal Processing for Storage
- Speech Recognition
- Storage Networks
- Tactical Communication
- Voice Over IP
- Voice XML

BUSINESS APPLICATION SESSIONS (BAS)

These will consist of presentations and panel discussions that generally provide a broader, more strategic, system-wide view of technology and markets than the technical sessions. Proposals are invited and must be submitted electronically, after 30 November 2002, to Terrie Heikkila (Email:terrie@donanderson.com). Deadline for Submission is 15 February 2003.

IMPORTANT DATES

Complete Paper Manuscripts Due: 15 February 2003
Paper manuscripts should be submitted electronically, after 30 November 2002, by using the Online Submissions form on the GLOBECOM 2003 website.

Proposals to Conduct Tutorials and Workshops Due: 15 February 2003
Proposals should be sent electronically, after 30 November 2002, to the Technical Program Chairs.

GENERAL CHAIR
Terry Kero (Myaani Inc.)
E-mail: tkero@myaani.com

TECHNICAL PROGRAM CHAIR
Prof. Willie W. Lu (CUHK / Siemens)
E-mail: wwlu@ieee.org

TECHNICAL PROGRAM VICE-CHAIR
Dr. Norival Figueira (Nortel Networks)
E-mail: norival@nortelnetworks.com

CONFERENCE MANAGEMENT
Terrie Heikkila, CMP (Don Anderson Inc.)
E-mail: terrie@donanderson.com

Our Push On Connectors Have Arrived

65GHz

High-Performance Low Cost

Anritsu Quality

Introducing Anritsu's New VP™ Connector Family

Now you can get Anritsu performance in a cost-effective Push On Connector solution. With throughput to **65 GHz** in a VP Bullet, VP Shroud, Cable Connector and VP-VF Adapters, the VP Connector's unique, small form factor design features:

- Hermetic connections • Sliding contact connection to microstrip • Testing capabilities
- Ground lip for handling substrates on carriers • Auto alignment capabilities

When you need Anritsu quality in a cost-effective Push On connector solution, **call 1-800-ANRITSU**, or check **www.us.anritsu.com/PushOn**

Anritsu

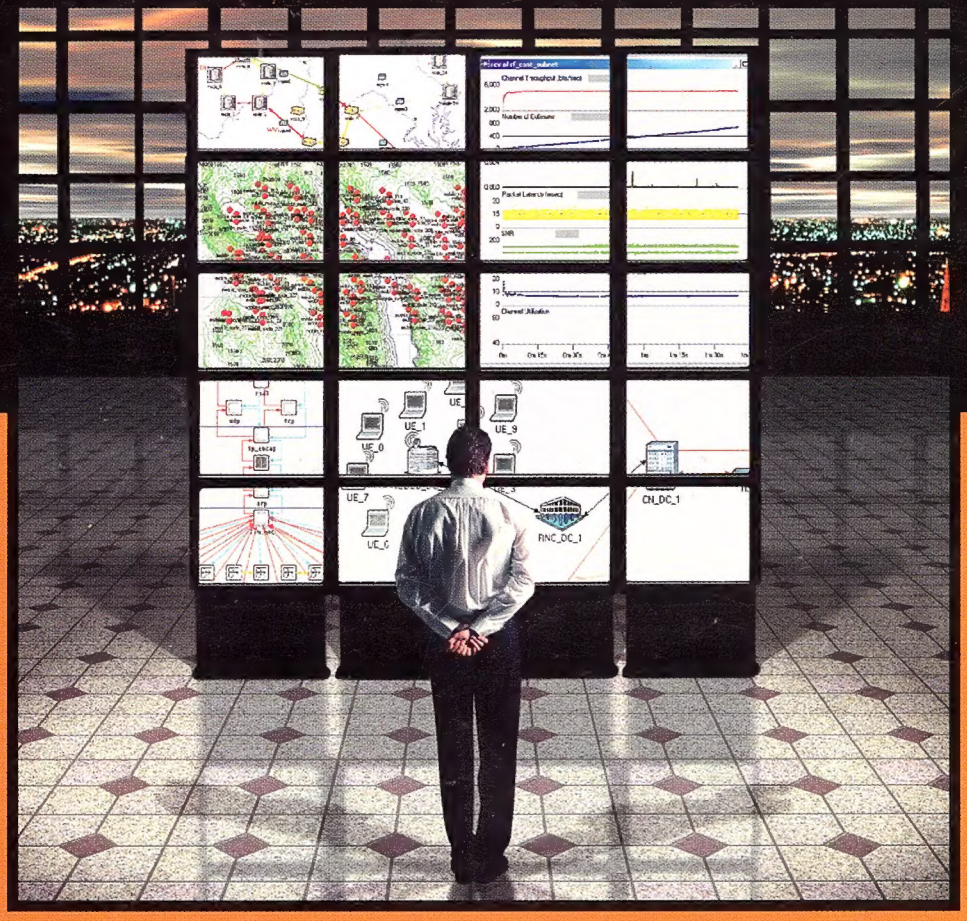
Discover What's Possible™

INTELLIGENCE TO...

VISIT OUR BOOTHS AT
3G World Summit
Tokyo, Japan
January 14-17, 2003

3GSM World Congress
Cannes, France
February 18-21, 2003
Booth #F4

... assess the efficiency of next generation wireless protocols
... accelerate the standardization of cutting edge technologies
... optimize network designs that incorporate proprietary technology
... study co-existence between UMTS and 802.11 newtorks
... analyze media content deployment on UMTS networks
... design QoS policies for your UMTS, ATM, and IP networks



Modeler
Accelerating Network R&D

IT Guru
Intelligent Network Management for Enterprises

SP Guru
Intelligent Network Management for Service Providers

Forbes 2002 BEST SMALL COMPANIES
OPNET Recognized for Financial Performance

Network Computing
COMPANY TO WATCH 2002

Winner of Supercom's **SUPERQUEST AWARD**

Founded in 1986, OPNET Technologies is the pioneer and leading provider of Intelligent Network Management software. OPNET software embeds expert knowledge about how network devices, network protocols, applications, and servers operate. This intelligence enables users in network operations, engineering, planning, and application development to be far more effective in optimizing performance and availability of their networks and applications. For more information about OPNET and its products, visit www.opnet.com.

***Software** That Understands Networks*

OPNET®

OPNET Technologies, Inc.
7255 Woodmont Avenue, Bethesda, Maryland 20814 • phone: (240) 497-3000 • fax: (240) 497-3001 • e-mail: info@opnet.com

www.opnet.com

Nasdaq: OPNT